Heterogeneous Credit Union Production Technologies with Endogenous Switching and Correlated Effects

Malikov, Emir and Restrepo-Tobon, Diego A and Kumbhakar, Subal C.

Auburn University, EAFIT University, State University of New York at Binghamton

May 2016

Heterogeneous Credit Union Production Technologies with Endogenous Switching and Correlated Effects*

Emir Malikov[1,2] Diego A. Restrepo-Tobón[3] Subal C. Kumbhakar[4,5]

[1]Department of Agricultural Economics & Rur. Soc., Auburn University, Auburn, AL
[2]Department of Economics, St. Lawrence University, Canton, NY
[3]Department of Finance, EAFIT University, Medellín, Colombia
[4]Department of Economics, State University of New York at Binghamton, Binghamton, NY
[5]University of Stavanger Business School, Stavanger, Norway

First Draft: April 15, 2013
This Draft: May 25, 2016

Abstract

Credit unions differ in the types of financial services they offer to their members. This paper explicitly models this observed heterogeneity using a generalized model of endogenous ordered switching. Our approach captures the endogenous choice that credit unions make when adding new products to their financial services mix. The model that we consider also allows for the dependence between unobserved effects and regressors in both the selection and outcome equations and can accommodate the presence of predetermined covariates in the model. We use this model to estimate returns to scale for U.S. retail credit unions from 1996 to 2011. We document strong evidence of persistent technological heterogeneity among credit unions offering different financial service mixes, which, if ignored, can produce quite misleading results. Employing our model, we find that credit unions of all types exhibit substantial economies of scale.

Keywords: Credit Unions, Correlated Effects, Ordered Choice, Panel Data, Production, Returns to Scale, Switching Regression

JEL Classification: C33, C34, D24, G21

* *Email*: emalikov@stlawu.edu (Malikov), drestr16@eafit.edu.co (Restrepo), kkar@binghamton.edu (Kumbhakar).

We would like to thank the editor, the associate editor and two anonymous referees for many insightful comments and suggestions that helped improve this article. We are also thankful to Alfonso Flores-Lagunes, Thierry Magnac, Dave Wheelock, Jeff Wooldridge and seminar participants at Syracuse University, University at Albany, Federal Reserve Bank of Dallas and the 2013 Midwest Econometrics Group Meeting at Indiana University Bloomington for many helpful comments and suggestions. Any remaining errors are our own. We also gratefully acknowledge the technical assistance from the Center for Scientific Computation APOLO at EAFIT University in setting up the cluster we used for the estimation. Restrepo-Tobón acknowledges financial support from the Colombian Administrative Department of Sciences, Technology and Innovation, Colombian Fulbright Commission and EAFIT University. The original version of this paper was circulated under the title "Are All U.S. Credit Unions Alike? A Generalized Model of Heterogeneous Technologies with Endogenous Switching and Correlated Effects".

1 Introduction

U.S. credit unions continue to prosper despite the decline in their relative advantages over commercial banks. Factors such as increasing availability of credit information from national credit-reporting bureaus, establishment of the federal deposit insurance fund for credit unions and the growth in credit card lending by larger financial institutions have significantly eroded conventional benefits of doing business at the local, small-scale level (Petersen and Rajan, 2002; Walter, 2006). This has motivated credit unions to evolve.

With the authorization to issue long-term mortgage loans in 1977 and the passage of the Credit Union Membership Access Act of 1998 which empowered them to widen and diversify their membership scope, credit unions have grown significantly in an attempt to compensate for the loss of traditional competitive advantages by capitalizing on economies of scale. Over the past decade, the average size of (federally-insured) credit unions has increased from $57.5 million to $135.8 million in assets. As of the end of 2011, the industry accounted for about a trillion dollars in assets and more than 92 million members (authors' calculations based on NCUA, 2011).

Several studies have investigated the performance of U.S. credit unions.[1] However, to our knowledge, no attempt has been made to formally model credit unions' technologies taking into consideration their differing output mixes (that is, different financial service menus they offer to their members). This limits our understanding of the industry structure, its evolution and the potential impact of alternative policies. Most previous studies have encountered the same problem, namely, the presence of a large number of observations for which the reported values of credit unions' outputs are zeros. This issue has been handled either by linearly aggregating different types of outputs into larger bundles (Fried et al., 1999; Frame and Coelli, 2001; Wheelock and Wilson, 2011, 2013) or by replacing zero outputs with an arbitrary small positive number (Frame et al., 2003). These methods may however be inappropriate since they do not recognize that the existence of zero-value outputs provides valuable information regarding the choice of the production technology by credit unions.

To preview the importance of modeling the choice of credit unions' technology (which we discuss in detail in Section 3), consider Table 1 which presents the number of retail credit unions in each year between 1994 and 2011 with zero values reported for some (or all) of the four outputs commonly considered in the literature. All credit unions[2] report non-zero values for consumer loans ($y3$) which historically have been their main product. However, there is a strikingly large number of credit unions that offer no real estate ($y1$) or business loans ($y2$) to their members throughout the years we consider. This evidence favors our view that not all credit unions are alike. Given that the output mix differs across units and over time, a substantial time-persistent heterogeneity may exist among credit unions.

We view this observed heterogeneity as an outcome of an endogenous choice made by credit union managers. They decide what range of services to offer to their members and choose the appropriate technology to provide them. Thus, it is likely that the production technology which a credit union employs varies with its output mix. To our knowledge, this technological heterogeneity (defined by the output mix) has been either assumed to be exogenous and/or entirely taken for granted in all previous studies. The aggregation of outputs into broader categories to solve the zero-output problem, so often practiced in the literature, constitutes the loss of information in both econometric and economic senses. The results previously reported in the literature are therefore likely to be misleading since the used econometric models ignore the time-persistent heterogeneity

[1]See Wheelock and Wilson (2011, 2013) and the references therein.
[2]With the exception of a single entity.

Table 1. Zero-Value Observations, 1994–2011

Year	$y1$	$y2$	$y3$	$y4$	Total	Year	$y1$	$y2$	$y3$	$y4$	Total
1994	3,670	9,063	0	3	9,783	**2004**	2,344	7,099	1	64	8,209
1995	3,517	9,056	0	0	9,734	**2005**	2,171	6,695	1	57	7,948
1996	3,555	9,162	0	2	9,891	**2006**	2,044	6,333	1	68	7,718
1997	3,441	9,059	0	0	9,765	**2007**	1,952	6,101	1	59	7,506
1998	3,269	8,811	0	0	9,561	**2008**	1,805	5,703	1	38	7,174
1999	3,140	8,650	0	55	9,426	**2009**	1,485	5,086	1	55	6,521
2000	2,925	8,442	0	75	9,195	**2010**	1,612	5,306	1	115	6,761
2001	2,764	8,114	0	61	8,932	**2011**	1,539	5,212	1	61	6,591
2002	2,601	7,739	0	61	8,611	**Total**	*46,377*	*133,152*	*9*	*870*	*151,817*
2003	2,543	7,521	1	96	8,491						

NOTES: The variables are defined as follows: $y1$ - real estate loans, $y2$ - business and agricultural loans, $y3$ - consumer loans, $y4$ - investments.

arising from the endogenous selection of credit unions' technologies.[3]

Heterogeneity among credit unions is unlikely to be limited to the technology they use; each credit union is unique in its operations. Ignoring this unobserved heterogeneity when estimating credit unions' technology (which is customary in the existing literature[4]) may produce inconsistent estimates since unobserved heterogeneity is likely to be correlated with covariates present in the estimated equation. While such credit-union-specific unobserved effects cannot be accounted for in a cross-sectional setting due to the incidental parameters problem, we address this issue in our case by taking advantage of the panel structure of the data.

In this paper, we address the above concerns by developing a unified framework that allows the estimation of credit union technologies that is robust to (i) misspecification due to an *a priori* assumption of homogeneous technology, (ii) selectivity bias due to ignoring the endogeneity in technology selection, and (iii) endogeneity (omitted variable) bias due to the failure to account for unobserved union-specific effects that are correlated with covariates in the estimated equations.

The estimation of such a model is not trivial. As we demonstrate in Section 3, the data indicate that 99% of all U.S. retail credit unions employ one of the three technologies associated with different output mixes offered by these institutions. These technologies have an *ordered* relationship: they go from a simpler to a more complex output mix. The existing literature on panel data selection models with unobserved heterogeneity focuses mainly on *binary* selection, and few papers allow for dependence between unobserved effects and covariates in *both* the outcome and the selection equations (see the references in Section 2). Among those studies that do allow for the latter, most rely on the assumption of strict exogeneity of covariates throughout the entire model (e.g., Wooldridge, 1995; Kyriazidou, 1997; Rochina-Barrachina, 1999) or at least in the selection equation (e.g., Charlier et al., 2001; Lee and Vella, 2006; Semykina and Wooldridge, 2010) which is particularly hard to justify in our application.

[3] Heterogeneity among credit unions has been also studied, although from a somewhat different perspective, in Wheelock and Wilson (2011) who estimate credit unions' cost function via kernel methods, thus avoiding any functional specification for the underlying technology and obtaining observation-specific estimates of the cost function. However, the aggregation of all types of loans into a single output, which the authors resort to, does not allow them to account for the (endogenous) heterogeneity resulting from differing output mixes, which our paper emphasizes.

[4] To our knowledge, Frame et al. (2003) is the only study which attempts to estimate (homogeneous) credit unions' technology using panel data while allowing for unobserved heterogeneity among institutions. However, the latter is modeled as random effects under a strong assumption of its exogeneity which is unlikely to be supported by the data.

Gayle and Viauroux (2007) study a dynamic panel data sample selection model quite similar to ours, where both the outcome and selection equations are permitted to contain predetermined variables as well as unobserved effects. However, to identify their model they require the presence of some strictly exogenous time-invariant variables in the selection equation. This assumption is however too restrictive for our application and is unlikely to be supported by the data. In a related study, Arellano and Carrasco (2003) study a single-equation binary choice dynamic panel data model with predetermined covariates and unobserved effects that are allowed to be correlated with the explanatory variables, which is similar to our technology selection equation. Given our empirical application, we propose a model of ordered selection, conditional on predetermined covariates, that allows for correlated unobserved effects in both the selection and outcome equations. To our knowledge, no such model has been considered in the literature.

We contribute to the literature by (i) extending Wooldridge's (1995) estimator to the case of ordered selection and the presence of predetermined covariates in the model and (ii) applying this framework to estimate the returns to scale for U.S. retail credit unions in 1996-2011. The latter has been recently brought into the spotlight of scholarly discourse (Emmons and Schmid, 1999; Wilcox, 2005, 2006; Wheelock and Wilson, 2011). We compare our estimates to those (potentially biased and inconsistent) obtained by ignoring heterogeneity due to endogenous technology selection and unobserved effects.

We find that not all U.S. retail credit unions are alike. There is evidence of persistent technological heterogeneity among credit unions offering different financial service mixes. We consistently reject the null hypotheses of exogenous technology selection and homogeneous (common) technology among credit unions. We further find that ignoring this observed heterogeneity and unobserved time-invariant effects across credit unions leads to downward biases in returns to scale estimates. In particular, models that do not account for parameter heterogeneity, endogenous switching and/or dependence between unobserved effects and right-hand-side covariates can produce the misleading finding that 6 to 20% of credit unions offering all types of loans suffer from diseconomies of scale and are thus scale-inefficient. This result broadly vanishes when we address all the concerns we raise in this paper. We find that most credit unions (of all technology types) exhibit substantial economies of scale. Hence, the growth of the industry is far from reaching its peak. We therefore expect a policy debate over credit unions' tax-exempt status and their special regulatory treatment compared with commercial banks to reignite in the near future. As these institutions grow in size and complexity, they may become of systemic importance. Regulators should be aware of these trends to contain threats that credit unions may potentially pose for local and national economies.

We also note that our generalized model is not tailored to the analysis of credit unions only. The framework can be applied to any other panel data study where selectivity and both observed and unobserved heterogeneity are present. Some examples would be studies of electric or water utilities, which often include both specialized and integrated companies that operate under non-homogeneous production technologies.

The rest of the paper proceeds as follows. We describe our panel data model of endogenous ordered switching under sequential exogeneity in Section 2. Section 3 provides a description of the data as well as a discussion of how we identify heterogeneous credit union technologies. Section 4 presents the results, and Section 5 concludes.

2 A Panel Data Model of Endogenous Ordered Switching

This section develops an econometric model that we employ to investigate underlying differences in heterogeneous technologies across U.S. credit unions. The model (i) avoids imposing a strong assumption of homogeneous technology uniformly adopted by all credit unions irrespective of the service mix they offer to their members; (ii) explicitly accounts for the endogeneity of the selection of these different technologies by unions over the course of time; and (iii) allows for unobserved time-invariant correlated effects amongst credit unions.

Consider a dual cost function of an s-type credit union i in period t:

$$C_{s,it} = \begin{cases} \mathbf{x}'_{s,it}\boldsymbol{\beta}_s + \alpha_{s,i} + u_{s,it} & \text{if} \quad T_{it} = s \\ - & \text{otherwise} \end{cases} \tag{2.1a}$$

$$T^*_{it} = \mathbf{d}'_{it-1}\boldsymbol{\rho}_t + \mathbf{z}'_{it-1}\boldsymbol{\gamma}_t + \xi_i + e_{it}\ , \quad (i = 1, \ldots, N;\ t = 1, \ldots, t_{max};\ s = 1, \ldots, S) \tag{2.1b}$$

where $C_{s,it}$ is the total variable, non-interest cost (the outcome variable); and $\mathbf{x}_{s,it}$ is a $K_s \times 1$ vector of strictly exogenous relevant cost function covariates as later defined in Section 3 (including unity for the intercept), with the corresponding parameter vector $\boldsymbol{\beta}_s$ of conformable dimension.[5]

The outcome variable $C_{s,it}$ is observed only if the sth technology is selected, i.e., if $T_{it} = s$. T^*_{it} is a latent variable governing the technology selection by a credit union i in period t, given the technology selected in the previous period T_{it-1} and an $L \times 1$ vector of some relevant lagged variables $\mathbf{z}_{it-1}$. We condition the technology selection in period t on the lagged technology T_{it-1} in order to allow for the state dependence of technology types over time. That is, a credit union naturally considers the financial services mix it currently offers to its members when making a decision about the composition of the mix for the next period. The state dependence is modeled via $\mathbf{d}'_{it-1}\boldsymbol{\rho}_t$, where $\mathbf{d}_{it-1} \equiv (\mathbb{1}\{T_{it-1} = 1\}, \ldots, \mathbb{1}\{T_{it-1} = S\})'$ with the corresponding $S \times 1$ parameter vector $\boldsymbol{\rho}_t$.[6] Further, we postulate the selection equation (2.1b) as a function of the lagged $\mathbf{z}$ variables in order to avoid making a strong assumption of contemporaneous or strict exogeneity of $\mathbf{z}$ which is unlikely to be supported by the data. Instead, we make a milder assumption of the predeterminedness of $\mathbf{z}_{it-1}$.[7] For a more elaborate discussion of the latter assumption, see Section 3. Parameter vector $(\boldsymbol{\rho}_t, \boldsymbol{\gamma}_t)$ is time-varying which allows for unrestricted temporal dynamics of e_{it}. Lastly, $(\alpha_{s,i}, \xi_i)$ are time-invariant, credit-union-specific unobserved effects. The subscript s denotes the technology type.

In Section 3, we show that the data point to three distinct credit union technologies associated with different output mixes and that there exists a clear ordered (nested) relationship between these technologies. Hence, it is natural to think of the latent variable T^*_{it} as measuring a credit union's propensity to select a more complex (diverse) output mix. The technology s is selected if and only if

$$T_{it} = s \quad \Leftrightarrow \quad \mu_{s-1,t} < T^*_{it} \leq \mu_{s,t}\ , \tag{2.2}$$

[5]In this paper, we consider the widely used translog cost function. Thus, to be exact, the left-hand-side variable will be the *log* of the total variable, non-interest cost, and the vector $\mathbf{x}_{s,it}$ will include the second-order log-polynomial of the cost function covariates. For more details, see Section 4.

[6]We thank an anonymous referee for suggesting this particular way of modeling state dependence. Clearly, for the identification purposes, one of the dummy variables for the past technology type needs to be dropped during the estimation.

[7]We acknowledge that the technology selection equation in (2.1b) may not be the true "structural" rule used by credit union managers when selecting their production technology. Rather, it can be thought of as a reduced-form representation of such a rule. Given that the technology selection process itself is of secondary interest, our choice of such a reduced-form modeling of switching seems reasonable.

where $\mu_{s,t} \in \{\mu_{0,t}, \ldots, \mu_{S,t}\}$ is a time-varying threshold.

Define $\mathbf{x}_{s,i} \equiv (\mathbf{x}'_{s,i1}, \ldots, \mathbf{x}'_{s,it_{max}})'$, $\mathbf{w}_{it} \equiv (\mathbf{d}'_{it}, \mathbf{z}'_{it})'$ and $\mathbf{w}_i^t \equiv (\mathbf{w}'_{i1}, \ldots, \mathbf{w}'_{it})'$. While we assume that the error terms $u_{s,it}$ and e_{it} are orthogonal to $\mathbf{x}_{s,i}$ and $\mathbf{w}_i^{t-1}$, their distributions are however allowed to be correlated, namely $\mathbb{E}\left[u_{s,it}e_{it} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}\right.\right] \neq 0$. Note that the above model is a generalization of a standard endogenous switching regression model to a case of ordered choice with the assumption of strict exogeneity of covariates in the selection equation being relaxed to weak (sequential) exogeneity.

The estimation of generalized model (2.1) is not trivial. While there has been a great interest in extending traditional limited dependent variable models to the case of panel data which permits controlling for unobserved effects, the literature on such models incorporated into linear regressions with selectivity mainly focuses on binary selection (for a comprehensive review, see Baltagi, 2013). These panel data selection models differ in their assumptions about the form of the unobserved heterogeneity in outcome and selection equations: whether (exogenous) random effects are assumed in both equations (e.g., Ridder, 1990; Verbeek and Nijman, 1996) or in the selection equation only (e.g., Verbeek, 1990). Few attempts have been made to allow for unobserved effects that correlate with right-hand-side covariates in both the outcome and selection equations. In the case of strictly exogenous covariates, some approaches to tackle such effects in panel sample selection models are those of Wooldridge (1995), Kyriazidou (1997) and Rochina-Barrachina (1999).[8] For a concise comparison of these estimators, see Dustmann and Rochina-Barrachina (2007).

Nonetheless, the above methods are not applicable in our case, since the selection equation (2.1b) contains predetermined covariates.[9] Gayle and Viauroux (2007) propose a three-stage semiparametric sieve estimator of a dynamic panel data sample selection model quite similar to ours in (2.1), where both the outcome and binary selection equations are permitted to contain predetermined variables as well as unobserved effects. However, one of the key restrictions needed to identify Gayle and Viauroux's (2007) model is the assumption that unobserved effects in the selection equation are correlated with a strictly exogenous time-invariant component of $\mathbf{z}_{it-1}$ only (in our notation). The latter assumption is however too restrictive for our application and is unlikely to be supported by the data, as discussed in Section 3. On the other hand, Arellano and Carrasco (2003) study a binary choice (dynamic) panel data model with predetermined covariates and unobserved effects that are allowed to be correlated with the explanatory variables, which is similar to our selection equation (2.1b).[10]

Given the research question that we posit in this paper, we consider a model of ordered choice, conditional on predetermined covariates, that allows for correlated unobserved effects in both the selection and outcome equations. To our knowledge, no such model has been considered in the literature. We thus fill this void by generalizing Wooldridge's (1995) estimator to the case of ordered selection and the presence of predetermined covariates in the model. For the generalization of Wooldridge (1995) to the case of polychotomous selection under strict exogeneity, see Malikov and Kumbhakar (2014).

We first formalize the selection equation (2.1b), where we build upon Arellano et al.'s (1999) and Arellano and Carrasco's (2003) setup.

Assumption 1. For $i = 1, \ldots, N$, $t = 1, \ldots, t_{max}$ and $s = 1, \ldots, S$:

[8] Also see Magnac (2000, 2004).

[9] Other similar panel data selection models relax strict exogeneity of covariates in the *outcome* equation (e.g., Charlier et al., 2001; Lee and Vella, 2006; Semykina and Wooldridge, 2010).

[10] The differences are: (i) we allow parameters to be time-varying, (ii) our selection process is not binary but ordered and, as we discuss later, (iii) we model correlated effects parametrically.

(i) The conditional mean of the unobserved effects in the selection equation is a linear projection on $\mathbf{w}_i^{t-1}$, i.e.,

$$\xi_i = \mathbb{L}\left[\xi_i \left| \mathbf{w}_i^{t-1}\right.\right] + c_i \tag{2.3a}$$

$$\mathbb{E}\left[c_i \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}\right.\right] = 0 \ , \tag{2.3b}$$

where $\mathbb{L}[\cdot]$ denotes the linear projection operator.

(ii) The composite error $\epsilon_{it} \equiv e_{it} + c_i$ is *i.i.d.* normally distributed over i given $(\mathbf{x}_{s,i}, \mathbf{w}_i^{t-1})$:

$$\epsilon_{it} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}\right. \sim \mathbb{N}\left(0, \sigma_t^2\right) \ . \tag{2.3c}$$

Thus, our model allows for dependence between unobserved effects ξ_i and right-hand-side covariates $\mathbf{w}_{it-1}$. Assumption 1 is slightly more restrictive than that in Arellano and Carrasco (2003, p.127) which we tighten by assuming the linearity of the conditional mean. In the latter respect, our approach is more close to that pursued by Arellano et al. (1999) who also model unobserved effects as a linear projection on the history of the predetermined covariates in their model. The benefit of assuming a linear conditional mean of ξ_i is that it allows us to dispense with a nonparametric estimation (via kernel methods) of conditional probabilities $\mathbb{P}\mathrm{r}\left[T_{it} = s \left| \mathbf{w}_i^{t-1}\right.\right]$, which one needs to do if following Arellano and Carrasco's (2003) approach. We seek to avoid a nonparametric estimation of the above probabilities primarily due to an acute "curse of dimensionality" problem associated with it which arises in our application given the high dimensionality of $\mathbf{w}_i^{t-1}$ (especially, when t approaches t_{max}).

Specifically, we let the linear projection $\mathbb{L}\left[\xi_i \left| \mathbf{w}_i^{t-1}\right.\right]$ in (2.3a) take the following form à la Mundlak (1978):

$$\mathbb{L}\left[\xi_i \left| \mathbf{w}_i^{t-1}\right.\right] = \overline{\mathbf{w}}_i^{t-1\prime} \boldsymbol{\eta}_t \ , \tag{2.4}$$

where $\overline{\mathbf{w}}_i^t = \frac{1}{\sum_{s=1}^t \mathbb{1}\{s \le t\}} \sum_{s=1}^t \mathbf{w}_i^s$, and $\boldsymbol{\eta}_t$ is an $(L+1)\times 1$ parameter vector. This is a quite popular parameterization of correlated effects in the literature (e.g., Semykina and Wooldridge, 2010). Alternatively, one can choose a less restrictive Chamberlain's (1980) specification $\mathbb{L}\left[\xi_i \left| \mathbf{w}_i^{t-1}\right.\right] = \mathbf{w}_i^{t-1\prime}\boldsymbol{\delta}_t$ which, for instance, underlines the selection process specified in Wooldridge (1995). Here, we opt for (2.4) due to its parsimony and relative computational simplicity.[11]

Under Assumption 1, the selection equation is given by

$$T_{it}^* = \mathbf{d}_{it-1}'\boldsymbol{\rho}_t + \mathbf{z}_{it-1}'\boldsymbol{\gamma}_t + \overline{\mathbf{w}}_i^{t-1\prime}\boldsymbol{\eta}_t + \epsilon_{it} \tag{2.5}$$

with the associated conditional probability of selecting the sth technology [in line with (2.2)]:

$$\begin{aligned} \mathbb{P}\mathrm{r}\left[T_{it} = s \left| \mathbf{w}_i^{t-1}\right.\right] = \Phi &\left(\frac{\mu_{s,t} - \mathbf{d}_{it-1}'\boldsymbol{\rho}_t - \mathbf{z}_{it-1}'\boldsymbol{\gamma}_t - \overline{\mathbf{w}}_i^{t-1\prime}\boldsymbol{\eta}_t}{\sigma_t}\right) - \\ &\Phi\left(\frac{\mu_{s-1,t} - \mathbf{d}_{it-1}'\boldsymbol{\rho}_t - \mathbf{z}_{it-1}'\boldsymbol{\gamma}_t - \overline{\mathbf{w}}_i^{t-1\prime}\boldsymbol{\eta}_t}{\sigma_t}\right) \ , \end{aligned} \tag{2.6}$$

where $\Phi(\cdot)$ denotes a standard normal cdf.

[11] In particular, Chamberlain's (1980) specification would require estimation of $[(L+1)t + (S-1)]$ parameters for each time period t. Due to high nonlinearity of the objective function and a relatively large t in our application, the true values of the parameters in (2.1a) may thus not be easy to locate.

Next, we formalize the treatment of unobserved effects in the outcome equation as well as the dependence between the two disturbances in (2.1a) and (2.5), where the latter enables us to correct for selection bias in the outcome equation.

Assumption 2. For $i = 1, \ldots, N$, $t = 1, \ldots, t_{max}$ and $s = 1, \ldots, S$:

(i) The conditional mean of the unobserved effects in the outcome equation s is a linear projection on $(\mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it})$, i.e.,

$$\mathbb{E}\left[\alpha_{s,i} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it}\right.\right] = \mathbb{L}\left[\alpha_{s,i} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it}\right.\right] \ . \tag{2.7a}$$

(ii) The error term $u_{s,it}$ is mean independent of $(\mathbf{x}_{s,i}, \mathbf{w}_i^{t-1})$ conditional on ϵ_{it} and is linear in ϵ_{it}, i.e.,

$$\mathbb{E}\left[u_{s,it} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it}\right.\right] = \mathbb{E}\left[u_{s,it} \left| \epsilon_{it}\right.\right] = \mathbb{L}\left[u_{s,it} \left| \epsilon_{it}\right.\right] \ . \tag{2.7b}$$

In particular, when modeling correlated effects in the outcome equation (2.1a), we consider the following general form of (2.7a) along the lines of Wooldridge (1995):

$$\begin{aligned}\mathbb{L}\left[\alpha_{s,i} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it}\right.\right] = {} & \mathbf{x}'_{s,i1}\boldsymbol{\varphi}_{s,t1} + \cdots + \mathbf{x}'_{s,it_{max}}\boldsymbol{\varphi}_{s,tt_{max}} + \\ & \mathbf{w}'_{i1}\boldsymbol{\omega}_{s,t1} + \cdots + \mathbf{w}'_{it-1}\boldsymbol{\omega}_{s,t(t-1)} + \psi_{s,t}\epsilon_{it} \ .\end{aligned} \tag{2.8}$$

Using the law of iterated expectations, one can easily show that, under our assumptions, the parameters on $\mathbf{x}_{s,it}$ in (2.8) are necessarily constant over t. Thus, (2.8) simplifies to

$$\begin{aligned}\mathbb{L}\left[\alpha_{s,i} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, \epsilon_{it}\right.\right] &= \mathbf{x}'_{s,i1}\boldsymbol{\varphi}_{s,1} + \cdots + \mathbf{x}'_{s,it_{max}}\boldsymbol{\varphi}_{s,t_{max}} + \\ &\quad \mathbf{w}'_{i1}\boldsymbol{\omega}_{s,t1} + \cdots + \mathbf{w}'_{it-1}\boldsymbol{\omega}_{s,t(t-1)} + \psi_{s,t}\epsilon_{it} \\ &= \mathbf{x}'_{s,i}\boldsymbol{\varphi}_s + \mathbf{w}_i^{t-1\prime}\boldsymbol{\omega}_{s,t} + \psi_{s,t}\epsilon_{it} \ ,\end{aligned} \tag{2.9}$$

where $\boldsymbol{\varphi}_s$, $\boldsymbol{\omega}_{s,t}$ and $\psi_{s,t}$ are $K_s t_{max} \times 1$, $(L+1)(t-1) \times 1$ parameter vectors and a scalar, respectively.[12] Note that this treatment of unobserved effects is in the spirit of Chamberlain (1980).

In Assumption 2(ii), the mean independence of $u_{s,it}$ in (2.7b) follows from the assumption of strict exogeneity of $\mathbf{x}_{s,it}$ and predeterminedness of $\mathbf{w}_{it-1}$ (as discussed in Section 3). Unlike Wooldridge (1995), we also condition the expectation of $u_{s,it}$ on $\mathbf{w}_i^{t-1}$. This is necessary because we allow the outcome and selection equations to have different covariates and non-zero (cross-equation) correlation between unobserved effects. Further, note that (2.7b) does not impose any restrictions on temporal dependence of $u_{s,it}$ or in the relationship between $u_{s,it}$ and ϵ_{it}.

Specifically, we set

$$\mathbb{L}\left[u_{s,it} \left| \epsilon_{it}\right.\right] = \pi_{s,t}\epsilon_{it} \ , \tag{2.10}$$

where parameter $\pi_{s,t}$ is allowed to be time-varying, thus emphasizing the presence of temporal dynamics in the relationship between $u_{s,it}$ and ϵ_{it}. The assumption of the disturbance in the outcome equation having a linear conditional mean is quite standard. A common alternative to it is the Heckman-type assumption of bivariate normality of the two disturbances which also implies linearity of the conditional mean of $u_{s,it}$. However, our assumption is less restrictive.

We are now ready to derive the selection bias corrected cost function. Taking the expectation of $C_{s,it}$ from (2.1a) conditional on the selection of the sth technology, we obtain

$$\mathbb{E}\left[C_{s,it} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, T_{it} = s\right.\right] = \mathbf{x}'_{s,it}\boldsymbol{\beta}_s + \mathbb{E}\left[\alpha_{s,i} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, T_{it} = s\right.\right] + \mathbb{E}\left[u_{s,it} \left| \mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, T_{it} = s\right.\right]$$

[12] Note that since $\mathbf{x}_{s,it}$ contains unity, the t_{max} intercept parameters in $\boldsymbol{\varphi}_s$ are not identified.

$$= \mathbf{x}'_{s,it}\boldsymbol{\beta}_s + \mathbf{x}'_{s,i}\boldsymbol{\varphi}_s + \mathbf{w}_i^{t-1\prime}\boldsymbol{\omega}_{s,t} + \varrho_{s,t}\mathbb{E}\left[\epsilon_{it}\left|\mathbf{x}_{s,i}, \mathbf{w}_i^{t-1}, T_{it} = s\right.\right]$$
$$= \mathbf{x}'_{s,it}\boldsymbol{\beta}_s + \mathbf{x}'_{s,i}\boldsymbol{\varphi}_s + \mathbf{w}_i^{t-1\prime}\boldsymbol{\omega}_{s,t} + \varrho_{s,t}\lambda_{s,it}\ , \tag{2.11}$$

where we have used (2.9) and (2.10) in the second equality. Here, $\varrho_{s,t} \equiv \pi_{s,t} + \psi_{s,t}$ and, given normality of ϵ_{it} under Assumption 1, $\lambda_{s,it}$ is the first moment of the truncated normal distribution.

Our generalized model is consistently estimated via a two-stage procedure. For each time period t, we first estimate the ordered probit via maximum likelihood as specified in (2.5)–(2.6). The parameter estimates of the selection equation are then used to obtain consistent estimates of $\lambda_{s,it}$. We then estimate the selection bias corrected cost function (2.11), in which predicted $\widehat{\lambda}_{s,it}$ are used in place of $\lambda_{s,it}$, via pooled least squares for each technology s, separately.

In order to conduct inference across equations for different technology types s as well as to account for the use of the predicted regressors $\widehat{\lambda}_{s,it}$ in the second stage, we follow Newey (1984) and cast the model in a multiple-equation system method-of-moments framework which permits derivation of an asymptotic variance-covariance matrix for our estimator. That is, by transforming the estimators from the two stages into their sample moment condition equivalents, i.e.,

$$f_N\left(\widehat{\boldsymbol{\theta}}\right) = \begin{bmatrix} \frac{1}{\sum \mathbb{1}\{t=1\}} \sum_i \frac{\partial \log \mathcal{L}_{i1}\left(\widehat{\boldsymbol{\theta}}_{\dagger,1}\right)}{\partial \widehat{\boldsymbol{\theta}}'_{\dagger,1}} \\ \vdots \\ \frac{1}{\sum \mathbb{1}\{t=t_{max}\}} \sum_i \frac{\partial \log \mathcal{L}_{it_{max}}\left(\widehat{\boldsymbol{\theta}}_{\dagger,t_{max}}\right)}{\partial \widehat{\boldsymbol{\theta}}'_{\dagger,t_{max}}} \\ \frac{1}{\sum \mathbb{1}\{T_{it}=1\}} \sum_i \sum_t \mathbf{h}_{1,it}\widehat{v}_{1,it}\left(\widehat{\boldsymbol{\theta}}_\dagger, \widehat{\boldsymbol{\theta}}_{\ddagger,1}\right) \\ \vdots \\ \frac{1}{\sum \mathbb{1}\{T_{it}=S\}} \sum_i \sum_t \mathbf{h}_{S,it}\widehat{v}_{S,it}\left(\widehat{\boldsymbol{\theta}}_\dagger, \widehat{\boldsymbol{\theta}}_{\ddagger,S}\right) \end{bmatrix}, \tag{2.12}$$

we can estimate the system-wide variance-covariance matrix by evaluating the (asymptotic) variance at our two-stage parameter estimates:

$$\left[\frac{\partial f_N(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}'}\right]^{-1} \widehat{\mathbf{V}}\left\{f_N\left(\widehat{\boldsymbol{\theta}}\right)\right\} \left[\left(\frac{\partial f_N(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}'}\right)'\right]^{-1}\ . \tag{2.13}$$

Here, $\widehat{\boldsymbol{\theta}}$ is an estimator of the system-wide parameter vector containing the first- and second-stage parameters. Specifically, the parameters of the first-stage technology selection equation are denoted by $\boldsymbol{\theta}_\dagger = (\boldsymbol{\theta}'_{\dagger,1}, \ldots, \boldsymbol{\theta}'_{\dagger,t_{max}})'$ with $\boldsymbol{\theta}_{\dagger,t} = (\mu_{1,t}, \ldots, \mu_{S,t}, \boldsymbol{\rho}'_t, \boldsymbol{\gamma}'_t, \boldsymbol{\eta}'_t, \sigma_t)'$, while the parameters of the second-stage selection bias corrected cost function are given by $\boldsymbol{\theta}_\ddagger = (\boldsymbol{\theta}'_{\ddagger,1}, \ldots, \boldsymbol{\theta}'_{\ddagger,S})'$, where $\boldsymbol{\theta}_{\ddagger,s} = (\boldsymbol{\beta}'_s, \boldsymbol{\varphi}'_s, \boldsymbol{\omega}'_s, \varrho_{s,3}, \ldots, \varrho_{s,t_{max}})'$. Further, $\log \mathcal{L}_{it}\left(\boldsymbol{\theta}_{\dagger,t}\right) = \sum_s \mathbb{1}\{T_{it} = s\} \log \mathbb{P}\mathrm{r}\left[T_{it} = s\left|\mathbf{w}_i^{t-1}\right.\right]$ is the log-likelihood function for the technology selection equation (2.5) for a credit union i in the time period t, with the probability $\mathbb{P}\mathrm{r}\left[T_{it} = s\left|\mathbf{w}_i^{t-1}\right.\right]$ given in (2.6). $\mathbf{h}_{s,it} = (\mathbf{x}'_{s,it}, \mathbf{x}'_{s,i}, \mathbf{w}_i^{t-1\prime}, \lambda_{s,it})'$ and $\widehat{v}_{s,it}$ are a column vector of right-hand-side covariates from the selection bias corrected cost function (2.11) and a corresponding least-squares residual for an s-type credit union i in the time period t, respectively. $\widehat{\mathbf{V}}$ is a robust estimate of the variance-covariance of the moment conditions.

Remark 1. Given that the selection equation includes lagged covariates, to estimate model (2.1) one needs to forgo the first wave of observations. Further, the parameterization of unobserved effects ξ_i in the selection equation as a linear projection on the time averages of $\mathbf{w}_i^{t-1}$ implies that $\overline{\mathbf{w}}_i^{t-1} = \mathbf{w}_i^{t-1} \equiv (\mathbf{d}'_{it-1}, \mathbf{z}'_{it-1})'$ for $t = 2$, resulting in perfect collinearity among right-hand-side variables in (2.5). The perfect collinearity arises because, for $t = 2$, the time averages of $\mathbf{w}_i^{t-1}$ are equal to $\mathbf{w}_i^{t-1}$. Specifically, $\overline{\mathbf{w}}_i^1 = \mathbf{w}_i^1$. Thus, when estimating our generalized model, one can effectively use observations for $t = 3, \ldots, t_{max}$ only.

Remark 2. Since the parameter vector in (2.11) has both time-invariant and time-varying components, we suggest organizing the data for each s and i as follows

$$\mathbb{E}\left(\left[\begin{array}{c} C_{s,i3} \\ \vdots \\ C_{s,it_{max}} \end{array}\right] \middle| \cdot \right) = \left[\begin{array}{c} \mathbf{x}'_{s,i3} \\ \vdots \\ \mathbf{x}'_{s,it_{max}} \end{array}\right] \boldsymbol{\beta}_s + \left[\begin{array}{c} \mathbf{x}'_{s,i} \\ \vdots \\ \mathbf{x}'_{s,i} \end{array}\right] \boldsymbol{\varphi}_s + \left[\begin{array}{ccc} \mathbf{w}_i^{2\prime} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \\ \mathbf{0} & & \mathbf{w}_i^{(t_{max}-1)\prime} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\omega}_{s,3} \\ \vdots \\ \boldsymbol{\omega}_{s,t_{max}} \end{array}\right] +$$
$$\left[\begin{array}{ccc} \lambda_{s,i3} & 0 & 0 \\ 0 & \ddots & \\ 0 & & \lambda_{s,it_{max}} \end{array}\right] \left[\begin{array}{c} \varrho_{s,3} \\ \vdots \\ \varrho_{s,t_{max}} \end{array}\right].$$

Note that the parameter vector $\boldsymbol{\omega}_s = (\boldsymbol{\omega}_{s,3}, \ldots, \boldsymbol{\omega}_{s,t_{max}})$ is $\frac{(t_{max}+1)(t_{max}-2)}{2}(L+1) \times 1$. In the case of a large t_{max}, equation (2.11) is likely to suffer from severe multicollinearity due to the inclusion of many $\mathbf{w}_i^{t-1}$ covariates. In such instances, we recommend restricting the elements of $\boldsymbol{\omega}_{s,t}$ to be equal, i.e., setting $\boldsymbol{\omega}_{s,t1} = \cdots = \boldsymbol{\omega}_{s,t(t-1)}$ in the notation used in (2.9). The latter implies that (2.11) ought to include $\overline{\mathbf{w}}_i^{t-1}$ in place of $\mathbf{w}_i^{t-1}$. This is equivalent to assuming that unobserved effects $\alpha_{s,i}$ take a (time-varying) Mundlak-type form in the $\mathbf{w}$ dimension [as opposed to Chamberlain's specification used in Assumption 2(i)]. This restriction significantly decreases the dimension of $\boldsymbol{\omega}_s$ to $(t_{max} - 2)(L+1) \times 1$.[13]

3 Heterogeneous Credit Union Technologies

Before we proceed, we note that the notation used in this section has no connection to that in previous sections unless specified otherwise.

3.1 Conceptual Framework

In this section, we define the framework in which we study credit union technologies. Due to their cooperative nature, credit unions are not profit-maximizers. Instead, they are thought of as maximizing service provision to their members in terms of quantity, price and variety of services (Smith, 1984; Fried et al., 1999). Following a wide practice in the literature (Frame and Coelli, 2001; Frame et al., 2003), we adopt a "service provision approach" under which, given their production technologies,[14] credit unions minimize variable, non-interest cost subject to the levels and types of outputs, the competitive prices of variable inputs and the levels of quasi-fixed netputs.

We consider the following four outputs: real estate loans ($y1$), business and agricultural loans ($y2$), consumer loans ($y3$) and investments ($y4$). We further follow Frame et al. (2003) and Wheelock and Wilson (2011, 2013) and include two quasi-fixed netputs (services) to capture the price dimension of the service provision by credit unions: the average interest rate on saving deposits ($\widetilde{y}5$) and the average interest rate on loans ($\widetilde{y}6$). The variable input prices that enter the credit union cost are the price of capital ($w1$) and the price of labor ($w2$). To partially account for the riskiness of the credit union operations, we also include equity capital ($\widetilde{k}$) as a quasi-fixed input in the cost

[13]We impose this restriction when estimating the model in Section 4.

[14]That is, given the mix of financial services (outputs) that credit unions opt to provide to their members.

function, as usually done in the banking literature. Credit unions studies have broadly ignored the latter under the implicit assumption of risk-neutral behavior of credit union managers. Including equity capital is also appropriate if one considers it as an additional input to the production of loans (e.g., see Hughes and Mester, 1998, 2013, among many others). These variables are taken as arguments of the dual (short-run) variable, non-interest cost function of a credit union, defined as

$$\mathcal{C}\left(\mathbf{y}, \widetilde{\mathbf{y}}, \mathbf{w}, \widetilde{k}\right) = \min_{\mathbf{x}} \left\{ \mathbf{x}'\mathbf{w} \mid T\left(\mathbf{y}, \widetilde{\mathbf{y}}, \mathbf{x}, \widetilde{k}\right) \leq 1;\ \widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}_0;\ \widetilde{k} = \widetilde{k}_0 \right\} , \tag{3.1}$$

where $\mathbf{y} = (y1, y2, y3, y4)$ is a vector of outputs, $\widetilde{\mathbf{y}} = (\widetilde{y}5, \widetilde{y}6)$ is a vector of quasi-fixed netputs with the corresponding vector of observed (fixed) values $\widetilde{\mathbf{y}}_0$; $\mathbf{w} = (w1, w2)$ is a vector of the variable input prices; $\mathbf{x} = (x1, x2)$ is a vector of variable inputs; $\widetilde{k}$ is a quasi-fixed input with the observed (fixed) value $\widetilde{k}_0$; and $T(\cdot)$ is the transformation function.

Compared to a primal specification of the production process, the dual cost approach is advantageous mainly because it avoids the use of input quantities which can lead to simultaneity problems given that the allocation of variable inputs is endogenous to a credit union manager's decisions. We thus treat the cost function covariates as strictly exogenous, as justified theoretically by the cost minimization premise and widely accepted in the financial services literature [e.g., see Hughes and Mester (2014) for an excellent review].[15]

The data we use in this study come from year-end call reports available from the National Credit Union Administration (NCUA), a federal regulatory body that supervises credit unions. The available data cover all state and federally chartered U.S. credit unions over the period from 1994 to 2011. We discard observations with negative values of outputs and total cost. Likewise, we exclude observations with non-positive values of variable input prices, quasi-fixed netputs, equity capital, total assets, reserves and total liabilities. Since $\widetilde{\mathbf{y}}$ and $w1$ are interest rates, we follow Wheelock and Wilson (2011) and also eliminate those observations for which values of these variables lie outside the unit interval. These excluded observations are likely to be the result of erroneous data reporting. For the details on construction of the variables from the call reports, see Appendix A.

In this paper we focus on retail, or so-called natural-person, credit unions only. We therefore exclude corporate credit unions (whose customers are the retail credit unions) from the sample to minimize noise in the data due to apparent non-homogeneity between these two types of depositories (this results in a loss of less than 0.7% of observations in the sample). Our data sample thus consists of 151,817 year-observations for all retail state and federally chartered credit unions over 1994–2011.

3.2 Heterogeneous Technologies

We next proceed to the identification of heterogeneous technologies among credit unions. As pointed out in the Introduction, the data indicate the presence of significant differences among credit unions in terms of the mix of services they offer to members. Based on the tabulation of zero-value observations reported in Table 1, on average, we find that 88% of credit unions in our sample do not offer business loans ($y2$) and 31% do not offer mortgage loans ($y1$) in a given year. Ignoring this observed heterogeneity in the provision of services across credit unions amounts to making a strong assumption that all credit unions share the same technology that is invariant to the

[15]Specifically, the outputs (loans) produced by banks and credit unions are normally said to be exogenous to these financial institutions because their quantities are determined by the customers' demand for loanable funds, which primarily depends on macroeconomic conditions beyond banks' control. Similarly, financial intermediaries are largely believed to be operating in perfectly competitive input markets which renders input prices exogenous to banks and credit unions.

Table 2. Tabulation of All Possible Heterogeneous Technologies, 1994–2011

Technology	Obs.	Unique CUs	Technology	Obs.	Unique CUs
Complete Specialization			*Three-Output Specialization*		
$y1$	5	1	$y1, y2, y3$	20	10
$y2$	0	0	$y1, y2, y4$	0	0
$y3$	673	328	$y1, y3, y4$	87,122	11,764
$y4$	0	0	$y2, y3, y4$	526	306
Two-Output Specialization			*No Specialization*		
$y1, y2$	0	0	$y1, y2, y3, y4$	18,118	4,466
$y1, y3$	171	113			
$y1, y4$	4	1			
$y2, y3$	1	1			
$y2, y4$	0	0			
$y3, y4$	45,177	9,446			

NOTES: The variables are defined as follows: $y1$ - real estate loans, $y2$ - business and agricultural loans, $y3$ - consumer loans, $y4$ - investments.

range of services they provide. This assumption is unlikely to hold since credit unions endogenously choose their output mixes.

Given the four types of loans we consider, we can identify 15 possible credit union technologies associated with unique output mixes. The possible heterogeneous technologies are those of the credit unions specialized in one (complete specialization), two or three types of loans (partial specialization) and of the unions that produce all four outputs (no specialization). Table 2 presents a summary of these technologies corresponding to output mixes constructed based on the non-zero-value loans reported by credit unions. The table shows that the majority of credit unions falls into the following three categories: (i) those that provide consumer loans and investments $\mathbf{y}_1 \equiv (y3, y4)$; (ii) those that provide real estate and consumer loans as well as investments $\mathbf{y}_2 \equiv (y1, y3, y4)$; and (iii) those that provide all types of outputs: real estate, business and consumer loans, and investments $\mathbf{y}_3 \equiv (y1, y2, y3, y4)$. Together, the three groups of credit unions constitute 99% of all observations in the sample, suggesting that the remaining one percent likely contains either outliers or reporting errors. We omit them from our analysis from this point forward. We label the three above output mixes as "1", "2" and "3", respectively, and define their corresponding technologies as "Technology 1", "Technology 2" and "Technology 3". We hereafter use technology and output mix types interchangeably when referring to credit unions. Also note that the three technology types are not independent but rather nested with a distinct ordering: a switch from Technology 1 (2) to Technology 2 (3) implies offering an extra output $y1$ ($y2$). Consequently, we model technology types as ordered alternatives in Section 2.

Figure 1 shows the breakdown of credit unions in our sample by the technology type. This figure indicates several trends. First, there is an apparent secular decline in the number of credit unions over time.[16] Second, the heterogeneity among U.S. credit unions (as captured by the technology type) is highly persistent. While today most credit unions still operate under Technology 2 as they did back in 1994, the presence of other technology types has increased over recent years. Third, there is a trend among credit unions to shift away from Technology 1 to Technology 2 and even more so to Technology 3 over time.

Table 3 presents summary statistics of the variables used in the dual cost function as well as

[16] Mainly due to mergers and acquisitions.

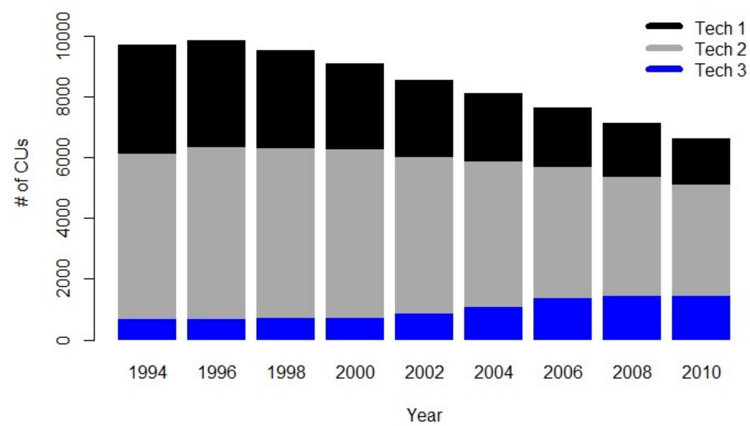

Figure 1. Tabulation of Credit Unions by Technology Type

several other variables descriptive of the characteristics of credit unions such as total assets, reserves, etc. All nominal stock variables are deflated to 2011 U.S. dollars using the GDP Implicit Price Deflator. A comparison of sample mean and median estimates of variables shows clear differences among credit union technologies. As expected, the size of the credit unions (proxied either by total assets, reserves or the number of members) increases as one moves from Technology 1 to Technology 3. This is also apparent in Figure 2 which plots kernel density estimates for the log of total assets tabulated by technology types. The large differences between technology types favor our view that the assumption of homogeneous (common) technology across credit unions is likely to result in the loss of information and the misspecification of the econometric model. As we show in Section 4, this produces biased estimates and potentially misleading results.

3.3 A Generalized Framework

We model the production technology for each of the three identified types of credit unions separately. We explicitly recognize that, under the abovementioned "service provision approach", credit unions minimize non-interest, variable cost subject to *different* types of outputs among other relevant constraints. Consequently, the associated production technologies are allowed to be heterogeneous over credit union types. That is, we consider the following generalization of the dual cost function (3.1)

$$\mathcal{C}_s\left(\mathbf{y}_s, \widetilde{\mathbf{y}}, \mathbf{w}, \widetilde{k}\right) = \min_{\mathbf{x}} \left\{\mathbf{x}'\mathbf{w} \mid T_s\left(\mathbf{y}_s, \widetilde{\mathbf{y}}, \mathbf{x}, \widetilde{k}\right) \leq 1;\ \widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}_0;\ \widetilde{k} = \widetilde{k}_0\right\} \qquad \forall\ s = 1, 2, 3\ , \tag{3.2}$$

where the output vector and the associated transformation and cost functions are indexed by one of the three types of credit unions s which we have identified above. Note that, unlike the model of homogeneous technology (3.1), the generalized model (3.2) does not suffer from the problem of having to deal with zero-value outputs.

Further, the above technological heterogeneity is likely to be an outcome of an endogenous choice made by credit unions. Based on the set of relevant demand and supply factors, credit union managers decide what range of financial services to offer to their members and choose the appropriate technology to provide them at the minimum cost. As seen above, the data particularly

Table 3. Summary Statistics, 1994–2011

Variable	Mean	Min	1st Qu.	Median	3rd Qu.	Max
			Technology 1			
Cost	171.8	0.7	47.6	101.2	205.3	9,866.0
$y3$	2,648.0	0.9	680.4	1,566.0	3,284.0	16,387.6
$y4$	1,547.0	0.0	167.9	580.3	1,635.0	262,500.0
$\widetilde{y}5$	0.028	0.000	0.017	0.029	0.038	0.056
$\widetilde{y}6$	0.100	0.000	0.082	0.095	0.110	0.993
$w1$	0.026	0.000	0.016	0.023	0.031	0.695
$w2$	32.9	0.0	20.1	32.2	43.3	266.3
$\widetilde{k}$	687.6	0.6	175.9	386.7	826.0	54,030.0
Total Assets	4,712.0	22.3	1,215.0	2,769.0	5721.0	373,600.0
Leverage	0.009	0.000	0.002	0.004	0.010	0.842
Reserves	198.8	0.0	47.6	100.2	214.0	18,270.0
Current Members #	1,127	27	401	745	1,378	43,560
Potential Members #	4,389	1	700	1461	3,000	10,000,000
Multiple-Bond CU	0.321					
Federal CU	0.625					
State CU (insured)	0.360					
			Technology 2			
Cost	2,244.0	3.2	333.4	767.5	1,965.0	580,500.0
$y1$	15,780.0	0.0	675.0	2,850.0	10,290.0	6,501,000.0
$y3$	24,750.0	3.0	3,767.0	8,172.0	20,090.0	9,126,000.0
$y4$	18,290.0	0.0	1,683.0	4,859.0	13,300.0	4,620,000.0
$\widetilde{y}5$	0.026	0.000	0.016	0.027	0.036	0.194
$\widetilde{y}6$	0.091	0.000	0.079	0.089	0.100	0.973
$w1$	0.026	0.000	0.016	0.023	0.031	0.695
$w2$	46.6	0.0	37.8	45.2	54.1	6,187.0
$\widetilde{k}$	7,338.0	0.8	1,080.0	2,477.0	5,955.0	2,587,000.0
Total Assets	65,750.0	116.0	8,908.0	20,580.0	51,300.0	24,090,000.0
Leverage	0.010	0.000	0.002	0.005	0.010	0.351
Reserves	2,638.0	0.0	294.7	707.5	1,800.0	2,563,000.0
Current Members #	8,859	5	1,754	3,570	8,276	2,451,000
Potential Members #	72,790	1	3,500	9,000	32,430	27,000,000
Multiple-Bond CU	0.427					
Federal CU	0.610					
State CU (insured)	0.378					

Table 3. Summary Statistics, 1994–2011 (cont.)

Variable	Mean	Min	1st Qu.	Median	3rd Qu.	Max
			Technology 3			
Cost	10,030.0	18.3	1,306.0	3,619.0	10,230.0	1,448,000.0
$y1$	119,400.0	1.0	8,314.0	29,230.0	94,810.0	18,940,000.0
$y2$	5,831.0	0.0	163.7	710.9	3,577.0	874,500.0
$y3$	98,490.0	13.0	10,260.0	29,440.0	84,190.0	14,340,000.0
$y4$	66,820.0	3.0	4,599.0	14,620.0	48,050.0	12,360,000.0
$\widetilde{y}5$	0.02	0.000	0.015	0.023	0.033	0.067
$\widetilde{y}6$	0.083	0.000	0.072	0.082	0.093	0.873
$w1$	0.026	0.000	0.016	0.023	0.031	0.695
$w2$	51.6	0.2	42.2	49.9	58.7	324.4
$\widetilde{k}$	32,970.0	10.0	3,902.0	10,250.0	29,870.0	5,079,000.0
Total Assets	326,400.0	224.0	35,860.0	98,320.0	288,600.0	46,930,000.0
Leverage	0.023	0.000	0.004	0.009	0.021	0.439
Reserves	11,880.0	0.0	1,106.0	2,956.0	8,159.0	4,906,000.0
Current Members #	32,070	119	4,972	12,570	33,070	3,867,000
Potential Members #	365,800	250	15,000	66,500	250,000	28,000,000
Multiple-Bond CU	0.307					
Federal CU	0.523					
State CU (insured)	0.457					

NOTES: The variables are defined as follows. *Cost* - total variable, non-interest cost; $y1$ - real estate loans, $y2$ - business and agricultural loans; $y3$ - consumer loans; $y4$ - investments; $\widetilde{y}5$ - average saving pricing; $\widetilde{y}6$ - average loan pricing; $w1$ - price of capital; $w2$ - price of labor; $\widetilde{k}$ - equity capital; *Leverage* - the ratio of total debt to total assets; *Multiple-Bond, Federal, and State (insured) CU* - indicator variables that take value of one if a CU is multiple-bond, federally accredited, or state-accredited (but federally insured), respectively. The remaining variables are self-descriptive. *Cost*, $y1$, $y2$, $y3$, $y4$, $w2$, $\widetilde{k}$, *Assets, Reserves* are in thousands of real 2011 US dollars; $\widetilde{y}5$, $\widetilde{y}6$, $w1$, *Leverage* are interest rates and thus are unit-free. The numbers of *Current and Potential Members* are in terms of number of people. Despite that minima of several variables are reported to be zeros (due to rounding), they are not exactly equal to zeros.

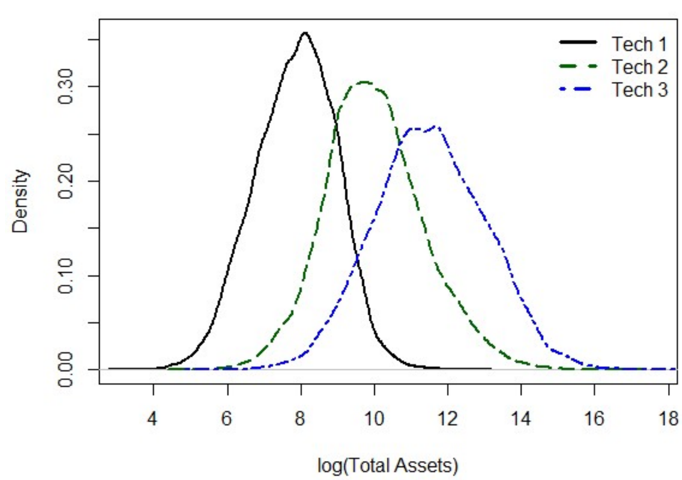

Figure 2. Kernel Densities of (log) Total Assets
Tabulated by Technology Type, 1994–2011

suggest considering covariates that correlate with the size of a credit union such as its total assets and other variables reflecting the credit union's financial strength and potential for growth and diversification. After carefully examining the existing literature for potential candidates, we settle on the following set of variables ($\mathbf{z}$): total assets, reserves, leverage ratio,[17] the number of current and potential members, indicator variables for federally accredited, state accredited and federally insured,[18] and multiple-bond credit unions. Table 3 provides their summary statistics.

We use the total value of assets and the number of current members of the credit union to capture the size of credit unions (Goddard et al., 2002). One can naturally expect a larger credit union to seek the diversification of its output mix and thus switch to a less specialized technology. We proxy the credit union's potential for growth using the reported level of reserves (Bauer, 2008; Bauer et al., 2009) and the size of the field of membership, i.e., the number of potential members (Goddard et al., 2008). The intuition here is as follows. The larger a credit union's field of membership is, the more likely it is to consider offering a wider range of services to its members and thus changing its technology. A larger membership field is likely to generate the demand for a more diverse menu of financial services. Similarly, the leverage ratio controls for the level of financial constraint a credit union may be subject to, which can directly influence its growth and the scope of services it offers. We also condition the choice of technology on whether a credit union can draw its members from a pool of people with single or multiple associations. This is crucial since multiple-bond credit unions have a substantial advantage over single-bond ones due to their ability to grow in size and diversify credit risks more easily (Walter, 2006). For instance, a single-bond credit union that is authorized to draw its members from a pool of employees of a single plant only is susceptible to any economic shock that this plant is subject to. Dummies for federally and state accredited credit unions are used to control for possible intrinsic differences between the two types of entities.

Unlike the cost function covariates, treating the above variables $\mathbf{z}$ as exogenous may however be invalid. While a larger credit union is able to offer a wider range of services to its members, the reverse may hold too: a more diversified credit union has a bigger capacity to grow. To avoid such an endogeneity problem when modeling technology selection, we conceptualize the output mix selection by credit unions as a lagged process. That is, we assume that a credit union considers its current position in terms of size, financial health, etc. as well as the service mix it currently offers to its members when making a decision about the composition of the mix for the next year. This seems reasonable given that a change in a credit union's service offerings is hardly an overnight venture but likely requires considerable time for activities like business planning and analysis, staff training, advertising, etc. Econometrically, the above assumption is equivalent to requiring that the *lagged* values of $\mathbf{z}$ be predetermined.

4 Estimation and Results

In order to analyze the consequences of the failure to accommodate heterogeneity in technologies resulting from endogenous selection as well as the presence of unobserved effects amongst credit unions, we estimate several auxiliary models in addition to the one developed in Section 2. For the ease of discussion, all the models we estimate are defined below.

Models Ignoring Unobserved Effects:

Model 1. The model of heterogeneous technologies with endogenous switching given by (2.1)

[17] Defined as the ratio of total debt to total assets.

[18] While all federally accredited unions are insured, the same however cannot be said about all state accredited unions.

where $\alpha_{s,i} = \xi_i = 0$. The model is estimated in two stages using (2.5)-(2.6) and (2.11) as described in Section 2, under the restriction $\boldsymbol{\eta}_t = \boldsymbol{\varphi}_s = \boldsymbol{\omega}_{s,t} = \mathbf{0}$.

Model 2. The model of homogeneous technology. This model is the most widely estimated in the literature by specifying two outputs instead of four in order to eliminate zero-value observations. The two outputs are the linearly aggregated loans ($y1 + y2 + y3$) and investments ($y4$). The model is estimated via pooled least squares using the whole sample ignoring a credit union's technology type.

Models Controlling for Unobserved Effects:

Model 3. The generalized model of heterogeneous technologies with endogenous switching and correlated effects given by (2.1) and estimated in two stages as described in Section 2. This is our preferred model.

Model 4. The model of homogeneous technologies with two outputs and correlated effects. The model is estimated via least squares using observations for credit unions of all technology types. In order to facilitate direct comparability between the models, here we model unobserved effects in the same fashion as in Model 3, i.e., by specifying the correlation between unobserved effects and the right-hand-side covariates in the spirit of Assumption 2(i).[19]

All models but generalized Model 3 are likely to be misspecified.[20] Further, note that Models 1 and 2 are special cases of Models 3 and 4, respectively. The sole difference between the two sets of models is that the correlated effects are assumed away in the first set (i.e., in Models 1 and 2). Comparing the results across these two sets of models enables us to gauge the degree to which the returns to scale estimates get distorted as a result of the potential model misspecification due to the ignored dependence between unobserved effects and covariates in the regressions.

Similarly, we estimate Models 2 and 4 to investigate how results change if one does not recognize technological heterogeneity among credit unions of different types. Both models are estimated under the most widely used specification in the literature, which assumes a common technology shared by all credit unions. Here, the misspecification is likely to stem from ignoring both the selectivity and heterogeneity in technologies. We assess the magnitude of distortions by comparing the estimates from Model 2 (4) with those from Model 1 (3).

For all models, we use the translog form[21] of the dual cost function, onto which we impose the symmetry and linear homogeneity (in input prices) restrictions. In the first stages of Models 1 and 3 (ordered probit), for the identification we suppress intercepts, normalize $\sigma_t = 1$ and set $\mu_{0,t} = -\infty$ and $\mu_{3,t} = \infty$. All continuous $\mathbf{z}$ variables that enter the selection equation are logged to allow for some degree of nonlinearity. To conserve space, we do not report the results from the first stage (they are available upon request) and thus directly proceed to the discussion of the main results.[22]

The left pane of Table 4 reports the summary statistics of the point estimates of returns to

[19]An alternative would be to estimate Model 4 via the within estimator that assumes no form of correlation between unobserved effects and covariates in the cost function (thus modeling these unobserved effects as "fixed effects").

[20]Provided our assumptions hold.

[21]While we emphasize the heterogeneity in credit unions' production technologies due to their differing output mixes, we acknowledge that ideally one would also prefer to allow the technology to be heterogeneous among credit unions for a *given* output mix. In this paper, we assume such heterogeneity away, which is an undeniable limitation of our analysis. One could extend our model to allow the cost function to be credit-union specific by, say, employing semi- or nonparametric methods (e.g., see Malikov et al., 2016), although controlling for unobserved effects in that case may require a different approach. Here, we opt for the parametric specification mainly for expository purposes as well as its tractability.

[22]The signs of statistically significant parameter estimates and mean marginal effects on conditional probabilities from the first-stage probit are all in line with the intuition.

Table 4. Summary of Returns to Scale Estimates

Model	Point Estimates of RS							Categories of RS, %		
	Mean	St. Dev.	Min	1st Qu.	Median	3rd Qu.	Max	DRS	CRS	IRS
Technology 1										
(1)	1.229	0.138	0.803	1.141	1.208	1.288	2.476	0.7	0.9	98.4
(2)	1.162	0.075	0.890	1.113	1.150	1.198	2.125	0.1	0.3	99.6
(3)	1.544	0.338	0.848	1.276	1.460	1.750	2.473	0.1	0.1	99.8
(4)	1.232	0.082	0.934	1.176	1.222	1.278	2.226	0.0	0.0	100.0
Technology 2										
(1)	1.084	0.059	0.876	1.042	1.080	1.120	1.799	5.0	2.9	92.1
(2)	1.085	0.065	0.878	1.040	1.078	1.120	2.162	5.0	2.6	92.4
(3)	1.372	0.259	0.914	1.167	1.315	1.533	2.488	0.5	0.4	99.1
(4)	1.149	0.087	0.922	1.089	1.137	1.193	2.189	0.6	0.7	98.6
Technology 3										
(1)	1.063	0.050	0.864	1.038	1.058	1.079	1.824	0.6	3.9	95.4
(2)	1.038	0.057	0.889	1.001	1.028	1.063	1.703	20.3	7.6	72.1
(3)	1.268	0.124	0.991	1.177	1.274	1.355	2.351	0.0	0.0	100.0
(4)	1.089	0.071	0.914	1.042	1.077	1.120	2.296	3.1	3.2	93.7
Whole Sample										
(1)	1.122	0.111	0.803	1.052	1.095	1.161	2.476	3.2	2.5	94.3
(2)	1.100	0.079	0.878	1.044	1.092	1.144	2.162	5.6	2.6	91.7
(3)	1.404	0.284	0.848	1.193	1.335	1.551	2.488	0.3	0.3	99.4
(4)	1.163	0.096	0.914	1.094	1.152	1.218	2.296	0.8	0.9	98.3

NOTE: Percentage points may not sum up to a hundred due to rounding.

scale based on all four models, over the 1996–2011 sample period.[23] Here, we break down the results by the technology type of credit unions. Note that although Models 1 and 3 estimate credit unions' cost functions for each technology separately, we also report the statistics for the whole distribution of credit unions obtained by pooling the results (over technology types) after the estimation. Similarly, we are able to break down the estimates of returns to scale from Models 2 and 4 by technology types after fitting a single homogeneous cost function for all credit unions. The credit-union-specific estimates of returns to scale are obtained using the formula that takes into account the quasi-fixity of equity capital (Caves et al., 1981)

$$RS = \frac{1 - \frac{\partial \log C}{\partial \log \widetilde{k}}}{\sum_j \frac{\partial \log C}{\partial \log y_j}} \ , \tag{4.1}$$

where $y_j \in \mathbf{y}_s$ are the outputs a credit union produces.

We first focus on the results from Models 1 and 2. The empirical evidence suggests that Model 2, which assumes a homogeneous production technology for all credit unions regardless of their differing output mixes, tends to underestimate the returns to scale for credit unions of Technologies 1 and 3, whereas the results are quite indistinguishable for Technology 2. Figure 3 shows these results by plotting kernel densities of the returns to scale estimates from all four models (for now, ignore those of the estimates from Models 3 and 4). We attribute this differences to biases in the

[23]The results are for 1996-2011 as opposed to 1994-2011 because the first two waves of the panel are consumed by lagged covariates and correlated effects in the technology selection equation as discussed in Section 2.

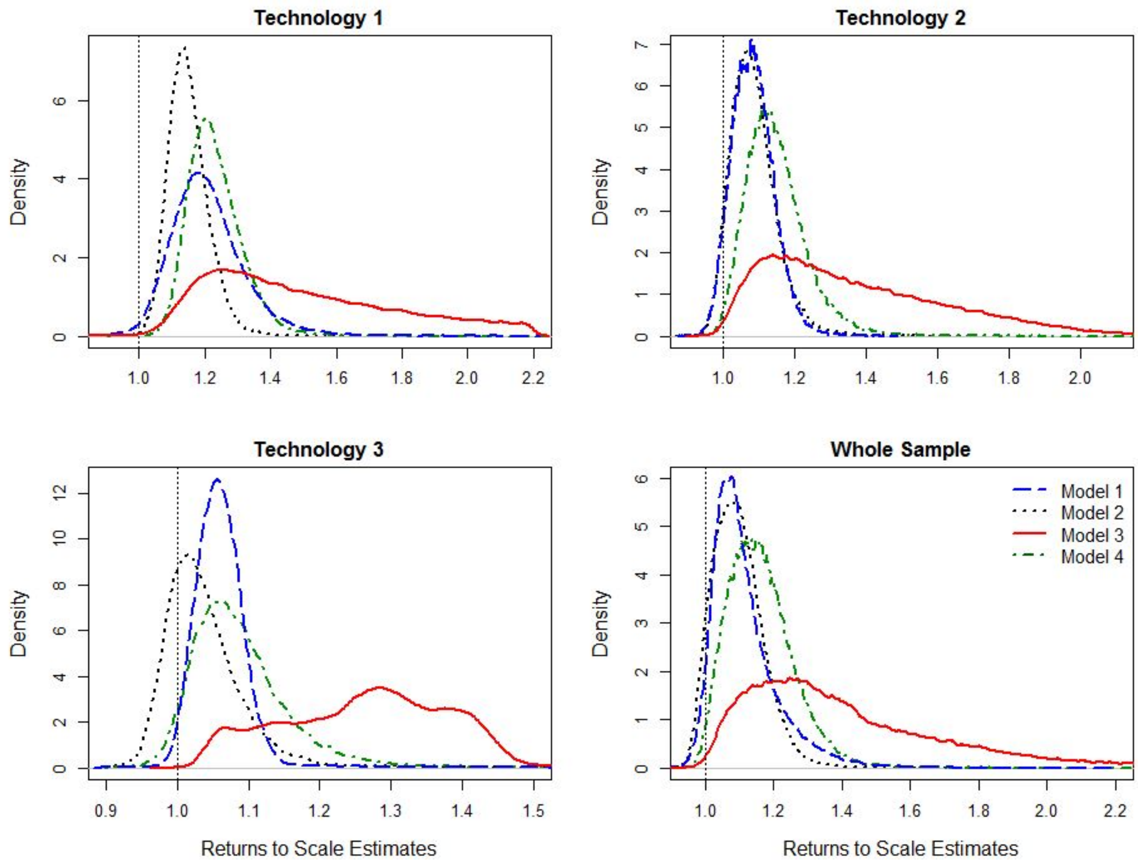

Figure 3. Kernel Densities of Returns to Scale Estimates

estimates from Model 2 due to the ignored selection and parameter heterogeneity.

We formally test the presence of non-homogeneous credit union technologies via the multiple-restriction Wald test of $\mathbb{H}_0: \ \boldsymbol{\beta}_s = \boldsymbol{\beta}_j$ for $s = 1, 2, 3$ $(s \neq j)$ in Model 1. The test strongly confirms the presence of heterogeneity in credit union cost structures: the p-value is less than 10^{-100}. We also perform a test for the presence of endogenous switching, i.e., a joint Wald test of $\mathbb{H}_0: \ \varrho_{s,3} = \cdots = \varrho_{s,t_{max}} = 0$ for $s = 1, 2, 3$ in Model 1. The tests reject the null of no selection bias with p-values less than 10^{-100} for all three technology groups, confirming that the switching is not exogenous and hence not "ignorable". The latter validates the proposition that the estimates from Model 2 are likely to be subject to selection and misspecification (due to imposed parameter homogeneity) biases.

The qualitative differences between the models are more transparent when credit unions are grouped into three returns to scale categories: decreasing returns to scale (DRS), constant return to scale (CRS) and increasing returns to scale (IRS). We classify a credit union as exhibiting DRS/CRS/IRS if the point estimate of its returns to scale is found to be statistically less than/equal to/greater than unity at the 95% significance level.[24] For clarity, we note that our notion of scale economies (i.e., increasing returns to scale) is defined as the cost savings that the credit union enjoys when increasing its scale of operation, with the cost per unit of output generally decreasing as fixed costs are spread over more units of output. In this paper, we only focus on such cost savings (if any)

[24] We use the delta method to construct standard errors for the returns to scale estimates.

within each credit union technology, as defined by the output mix. That is, our returns to scale measure does not include potential cost savings associated with the provision of a more diversified output mix if the credit union switches from Technology 1 to Technology 2 or to Technology 3. Such cost savings *across* credit union technologies would have rather been indicative of scope economies, which is a matter of substantial interest on its own and deserves a close examination in a separate paper.

Based on the results from Model 1 (see Table 4), we find that virtually all credit unions of Technology 1 operate under IRS. We however cannot say the same with respect to credit unions of the other two technology types. Here we find that 7.9% and 4.5% of credit-union-years under Technology 2 and 3 exhibit non-IRS (i.e., DRS or CRS), respectively. Qualitatively, Models 1 and 2 produce similar results for credit unions operating under Technology 1 and 2. However, the biases in estimates from Model 2 tell a rather different story for Technology 3. According to this model, astounding 20.3% of credit unions operate under DRS and are thus scale-inefficient.

However, as mentioned above, Models 1 and 2 are likely to be misspecified and their results may be misleading because of endogeneity bias due to the ignored dependence between unobserved effects and covariates in the regressions. We thus proceed to the models that explicitly control for correlated effects: Models 3 and 4.[25] Figure 3 plots the kernel densities of the returns to scale estimates from these models (see Table 4 for the summary statistics of the estimates).

The evidence suggests that the model which ignores endogenous switching and technological heterogeneity (Model 4) tends to underestimate the returns to scale at which credit unions operate across all three technology groups. The kernel densities of estimates from Model 3 are generally shifted rightward compared to those of estimates from Model 4. Thus, the biases in returns to scale estimates produced by Model 4 generally appear to be of negative sign.

We again reject the null of a homogeneous (common) cost function across different technology groups. The p-value corresponding to the Wald test of $\mathbb{H}_0: \ \boldsymbol{\beta}_s = \boldsymbol{\beta}_j$ for $s = 1, 2, 3$ $(s \neq j)$ on the coefficients of (2.11) in Model 3 is less than 10^{-100}. Similarly, the Wald tests of $\mathbb{H}_0: \ \varrho_{s,3} = \cdots = \varrho_{s,t_{max}} = 0$ for $s = 1, 2, 3$ performed on (2.11) again confirm the presence of selection bias in Model 4 (p-values are less than 10^{-100} for all three technology groups).[26] Thus, the data favor our preferred generalized Model 3.

Figure 3 also informs of the differences across Models 3 and 4, which account for credit union-specific correlated effects, and Models 1 and 2, which ignore this unobserved heterogeneity. The evidence indicates the presence of a negative bias in the returns to scales estimates obtained from Models 1 and 2: the kernel densities from these models are to the left of those produced by the corresponding models that control for unobserved effects. The biases appear to be the largest in the case of Technology 3. The above emphasizes the importance of taking unobserved effects into account when estimating credit union technologies.

Figure 4 depicts the 95% confidence intervals of the returns to scale estimates from generalized

[25] Following equation (2.11), we parameterize correlated effects in cost functions as linear projections of (i) all continuous variables included in the first-stage selection equation and (ii) all unique variables in the cost functions, except for the time trend. Thus, we do not include squared and cross-product terms from the translog cost functions into the set of variables onto which unobserved effects are assumed to project. Doing the latter would be redundant.

[26] Note that our assumptions imply two potential channels for selection bias in the outcome equation of interest: (i) the potential dependence between correlated effects in the outcome equation $\alpha_{s,i}$ and the error in the selection equation ϵ_{it} and (ii) the potential dependence between the error in the outcome equation $u_{s,it}$ and ϵ_{it}. These two dependencies are regulated by $\psi_{s,t}$ in (2.9) and $\pi_{s,t}$ in (2.10), respectively, the sum of which is defined to be $\varrho_{s,t}$ in the selection bias corrected equation (2.11). Hence, while we can use the joint exclusion Wald test on the $\varrho_{s,t}$ parameters to test for exogeneity/endogeneity of the selection, we are generally unable to formally discriminate between either of its two channels.

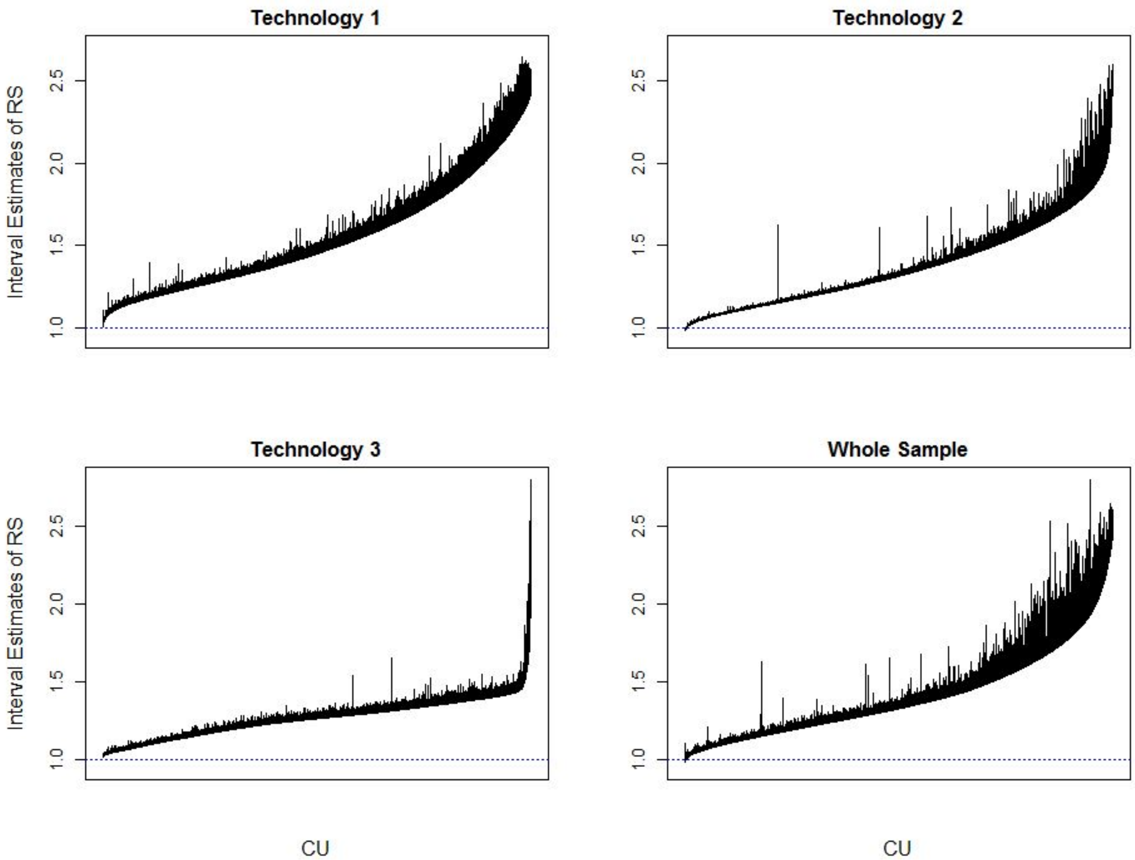

Figure 4. The 95% Confidence Intervals of Returns to Scale Estimates from Generalized Model 3

Model 3, based on which the right pane of Table 4 is partly populated. These confidence intervals, which correspond to each observation (credit-union-year) over the 1996–2011 period, are represented by vertical line segments that are sorted by the lower bound.[27] As expected, in contrast to Model 1, which ignores unobserved effects, Model 3 predicts virtually zero credit unions with non-IRS across all technology groups: virtually all confidence intervals lie above unity. In contrast, the results from Model 4 of homogeneous technology still suggest that 6.3% of credit unions of the third technology type exhibit non-IRS (see Table 4). The latter finding however is not as drastic as the one based on Model 2, a correlated-effects-free counterpart of Model 4.

Although both Models 3 and 4 strongly support the evidence in favor of IRS almost universally exhibited by credit unions operating under Technologies 1 and 2, the correspondence in rankings of credit unions by these models is weak. The Spearman's rank correlation coefficient of the returns to scale estimates from the two models is between 0.65 and 0.79. We attribute these differences to selection and misspecification biases present in Model 4.

As briefly mentioned above, we find the least agreement in results across our generalized Model 3 and Model 4 in the case of Technology 3: the rank correlation coefficient is 0.21. While Model 4 indicates that 3.1% and 3.2% of credit unions in this technology group operate at DRS and

[27]Note that the credit-union-years are not sorted by their respective RS point estimate but rather by the lower bound of the corresponding 95% confidence interval. The plot therefore contains "spikes" because the RS estimates for different credit-union-years have different confidence intervals, some of which are larger/smaller than others.

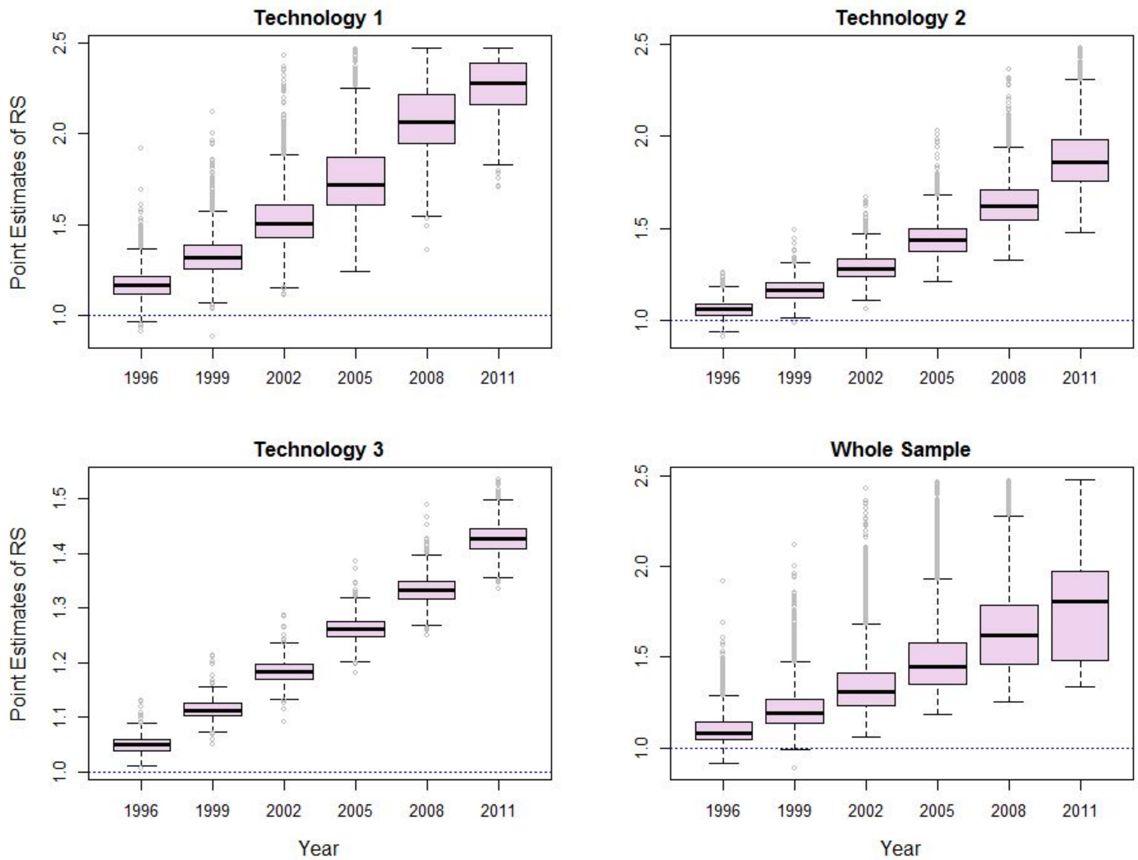

Figure 5. Returns to Scale over Time; Estimates from Generalized Model 3

CRS, respectively, based on our preferred Model 3 we however find that virtually all of these credit unions enjoy IRS. Incidentally, this finding is consistent with the results in Wheelock and Wilson (2011) who find no significant evidence of DRS and CRS among credit unions in their sample either. However, it is worth pointing out that, despite qualitative similarities between Wheelock and Wilson's and our findings (based on generalized Model 3), the results are not directly comparable. See Appendix B for a more elaborate discussion.

We find that returns to scale in the credit union industry have increased over the course of years, as can be seen in Figure 5. The phenomenon is observed for *all* technology types of credit unions. However, we find unexpected results when analyzing the relationship between returns to scale of a credit union and its size (proxied by total assets). Normally, one would expect to see an inverse relationship between the two. We do confirm it when looking at the entire sample. However, as Figure 6 shows, this result is not uniform across all technology groups. We find that the estimated returns to scale (from our generalized Model 3) fall as one moves from small to larger credit unions that operate under Technologies 1 and 2. However, the returns to scale increase with the size for credit unions operating under Technology 3. For instance, the estimates of returns to scale from Model 4 fall with the asset size *regardless* of the technology type (not reported to conserve space). While this finding looks puzzling at the first glance, there is an intuitive explanation to it.

Recall that the asset size of the credit unions increases as one moves from Technology 1 to 3 (see Table 3 and Figure 2). Thus, as credit unions grow and transition from the first technology type

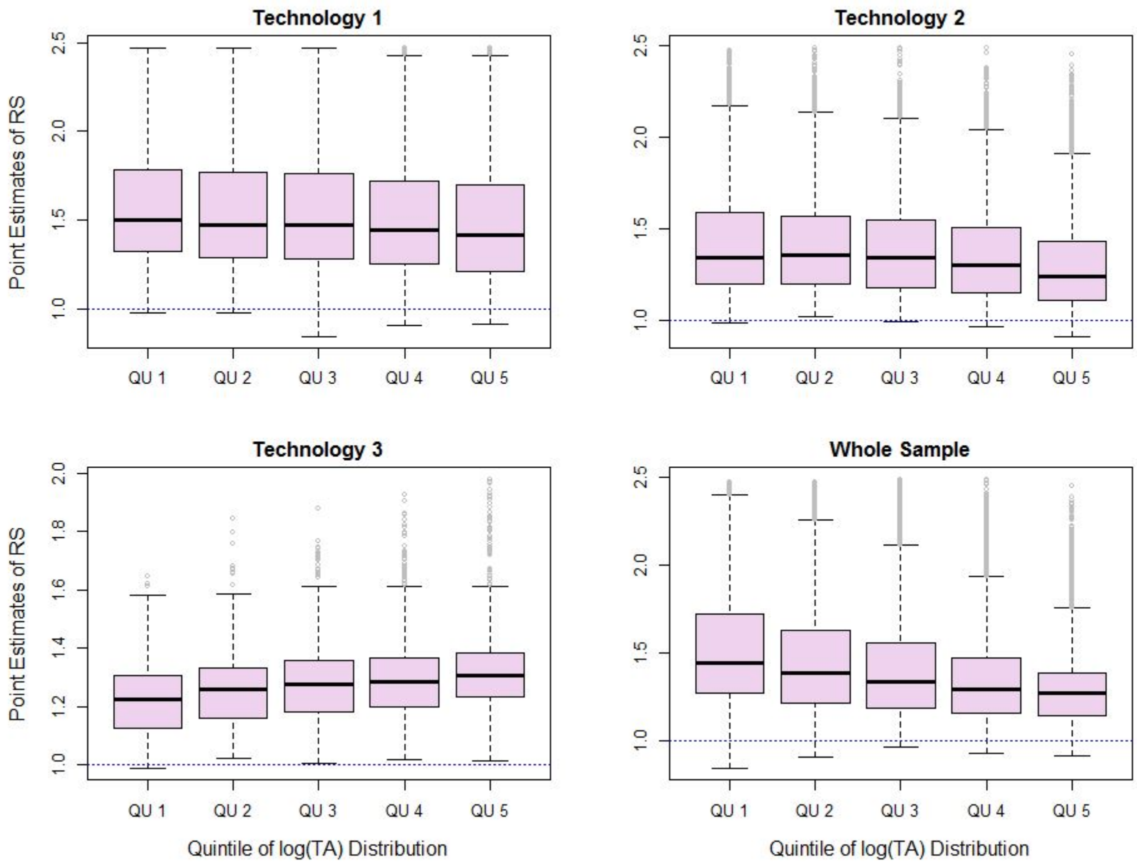

Figure 6. Returns to Scale by (log) Total Assets Quintiles; Estimates from Generalized Model 3

to the second, a positive effect of scale on the cost naturally wears out. The relationship between the size and returns to scale however breaks down for credit unions in the third technology group. One can think of several reasons to explain this. First, an increase in available resources as credit unions continue to grow enables them to adopt new information processing technologies that are unaffordable to smaller, more financially constrained credit unions but are substantial cost-savers. The example of such technologies would be internet banking, automated teller machines, use of electronic money as well as an access to members' credit history through the credit rating bureaus. Second, larger credit unions enjoy greater diversification. On average, credit unions in this group have a 32 (4) times larger number of members than those belonging to the first (second) technology group. The diversification comes not only through a larger membership pool, but also through a wider range of services provided to members as well as an opportunity to engage in more advanced financial operations (Wilcox, 2005). The latter is partly due to economies of diversification enjoyed by credit unions as they move from one technology to another (recall that technologies are ordered). The data suggest the presence of non-negligible economies of scope, which is a matter of substantial interest on its own. We leave the discussion of it for a future paper. Lastly, larger credit unions can also protect their market positions by erecting entry barriers thus partly mitigating the decline in returns to scale as they grow. Hughes and Mester (2013) report a similar finding for large banks.

5 Conclusion

A trillion dollar worth credit union industry takes up a significant portion of the U.S. financial services market, catering to almost a hundred million people in the country. Given the dramatic growth of the industry over the past few decades, there has been a substantial interest in formally modeling the technologies of credit unions. However, the econometric approaches widely used in the existing literature somewhat limit our understanding of the structure, dynamics and future evolution of the credit union industry.

Faced by the presence of an overwhelming number of observations for which the reported values of credit unions' outputs are zeros, the existing studies of credit union technologies have mainly resorted to the linear aggregation of different types of outputs into broader categories. This procedure leads to a loss of valuable information in both econometric and economic senses. In this paper, we show that the presence of zero-value observations is not merely a data issue but a consequence of substantial time-persistent heterogeneity amongst credit unions' technologies as captured by differing output mixes. This heterogeneity is likely to be an outcome of an endogenous choice made by credit unions. Models that *a priori* impose homogeneity and/or overlook credit unions' endogenous technology selection are likely to produce biased, inconsistent and misleading estimates. The results are also likely to be biased due to unobserved effects which are widely ignored in the credit union literature.

We address the above concerns by developing a unified framework that allows the estimation of credit union technologies that is robust to (i) misspecification due to an *a priori* assumption of homogeneous technology, (ii) selectivity bias due to ignoring the endogeneity in technology selection, and (iii) endogeneity (omitted variable) bias due to a failure to account for unobserved union-specific effects that are correlated with covariates in the estimated equations.

We develop a generalized model of endogenous switching with ordered choice and correlated effects that allows treatment of predetermined variables in the selection equation by building on Wooldridge's (1995) estimator. We note that our model is not tailored to the analysis of credit unions only. The framework can be applied to any other panel data study where selectivity and both observed and unobserved heterogeneity are present. Some examples would be studies of electric or water utilities, which often include both specialized and integrated companies that operate under non-homogeneous technologies.

We find that not all U.S. retail credit unions are alike. There is evidence of persistent technological heterogeneity among credit unions offering different financial service mixes. We consistently reject the null hypotheses of exogenous technology selection and homogeneous technology among credit unions and generally find that ignoring this observed heterogeneity or ignoring unobserved time-invariant effects across units leads to downward biases in returns to scale estimates.

Appendix A

Table A.1. Call Report Definitions of the Variables

Variable	NCUA Account Definition	Description
$y1$	Acct_703 + Acct_386	Real estate loans: first mortgage real estate loans, other real estate loans
$y2$	Acct_475	Commercial loans: business and agricultural loans (MBLs) granted YTD
$y3$	Acct_025B – $y1$ – $y2$	Consumer loans: total loans, less real estate loans, less commercial loans
$y4$	Acct_799	Total investments
$\widetilde{y}5$	(Acct_380 + Acct_381)/ Acct_018	Average interest rate on saving deposits: dividends on shares, interest on deposits, divided by total shares and deposits
$\widetilde{y}6$	(Acct_110 + Acct_131)/ Acct_025B	Average interest rate on loans: total (gross) interest and fee income on loans, fee income, divided by total loan and leases
$w1$	(Acct_230 + Acct_250 + Acct_260 + Acct_270 + Acct_280 + Acct_290 + Acct_310 + Acct_320 + Acct_360)/Acct_018	Price of capital: travel and conference expense, office occupancy expense, office operations expense, educational and promotional expense, loan servicing expense, professional and outside services, member insurance, operating fees (examination and/or supervision fees), miscellaneous operating expenses, divided by total shares and deposits
$w2$	Acct_210/(Acct_564A + 0.5*Acct_564B)	Price of labor: employee compensation and benefits, divided by full-time equivalent employees [Number of credit union employees who are: Full-time (26 hours or more) + 0.5*Part-time (25 hours or less per week)]
$\widetilde{k}$	Acct_931 + Acct_668 + Acct_945 + Acct_658 + Acct_940 + Acct_602	Equity: regular reserves, appropriation for non-conforming investments, accumulated unrealized gains (losses) on available-for-sale securities and other comprehensive income, other reserves, undivided earnings, net income
C	Acct_010	Total variable, non-interest cost: total non-interest expenses
Total Assets	Acct_010	Total assets
Leverage	(Acct_860C + Aacct_820a + Acct_825 + Acct_018)/ Acct_010	Total liabilities [total borrowing, accrued dividends and interest payable on shares and deposits, accounts payable and other liabilities, total shares and deposits], divided by total assets
Reserves	Acct_931 + Acct_668	Regular reserves, appropriation for non-conforming investments
Current Members #	Acct_083	Total number of current members
Potential Members #	Acct_084	Total number of potential members

Appendix B

As discussed in Section 4, based on our preferred Model 3, which accounts for both the endogenous choice of the output mix as well as correlated unobserved heterogeneity, we find that virtually all U.S. credit unions (99.4% of the sample) enjoy IRS during our sample period. Incidentally, this finding is consistent with the results in Wheelock and Wilson (2011) who also find significant evidence of IRS among credit unions in the U.S. However, despite qualitative similarities between Wheelock and Wilson's and our findings, the results are not directly comparable.

First, our sample periods differ: we consider the period of 1994–2011, whereas Wheelock and Wilson (2011) examine the 1989–2006 period.[28] Second, Wheelock and Wilson (2011) obtain their returns to scale estimates from an admittedly more flexible nonparametric cost function whereas our estimation approach is parametric. Third, they aggregate outputs in order to eliminate zero-value observations and fit a homogeneous production technology for all credit unions. Their cost function also does not include equity capital as one of the inputs. Fourth, Wheelock and Wilson (2011) do not explore the possibility of endogeneity in a credit union's choice of the output mix. Lastly, while controlling for time effects, Wheelock and Wilson (2011) however left the issue of unobserved time-invariant effects unaddressed. All of these issues undercut the comparability of Wheelock and Wilson's (2011) and our results.

In order to take a closer look at our results, we use our data sample to estimate an auxiliary model that closely follows Wheelock and Wilson's (2011) approach. To at least partly mitigate the above-referenced comparability problem, we adapt Wheelock and Wilson's (2011) nonparametric specification of a homogeneous credit union technology to this paper's setup with some fine-tuning. Specifically, the point of our departure is Wheelock and Wilson's (2011) baseline model given in equation (4) of their paper. To ensure a more meaningful comparison of our results with those obtained via their approach, consistent with our specification of the production technology of credit unions, we additionally condition the credit union's cost function on a quasi-fixed equity capital. The stochastic cost function is then defined as follows:

$$C = \mathcal{C}\big(\mathbf{y}, \widetilde{\mathbf{y}}, \mathbf{w}, \widetilde{k}\big) + u,$$

where $\mathcal{C}(\cdot)$ is an unspecified nonparametric function, $\mathbf{y} \equiv \big[(y1 + y2 + y3), y4\big]$, and u is a stochastic error. As earlier, we impose the linear homogeneity property by dividing the cost C and all input prices $\mathbf{w}$ by the price of labor ($w2$).

Following Wheelock and Wilson (2011), the model is estimated in logs via a local-linear fitting. Unlike them, we however do not employ any dimension reduction techniques, the usual argument for the use of which is the mitigation of the so-called "curse-of-dimensionality" problem whereby the nonparametric estimator's rate of convergence decreases with the number of continuous covariates. Given the fairly large sample size ($> 150,000$ observations) and a rather small number of continuous regressors in the model (a total of six), the "curse of dimensionality" is unlikely to cause severe problems in our application. More importantly, the principal component extraction technique favored by Wheelock and Wilson (2011) for reducing the dimensionality of data may not be a suitable solution here because it relies on the independent *linear* information in the data. Given that the core purpose of a nonparametric specification is to allow for potential *non*linearities in the conditional mean of the cost function, the use of principal components therefore appears to be rather self-defying.

Table B.1 reports the median nonparametric estimates of the expansion path scale economies (EPSE), a preferred measure of returns to scale for credit unions by Wheelock and Wilson (2011). EPSE is computed via the formula given in equation (11) in Wheelock and Wilson (2011):

$$EPSE = \frac{\mathcal{C}\big((1+\gamma)\mathbf{y}, \cdot\big)}{\left(\frac{1+\gamma}{1-\gamma}\right)\mathcal{C}\big((1-\gamma)\mathbf{y}, \cdot\big)},$$

where, following the authors, we set $\gamma = 0.05$. That is, the reported EPSE measures the returns to scale exhibited by a credit union along its expansion path going from the origin through its observed

[28] Unfortunately, we have no access to public data on credit unions that date back beyond 1994.

Table B.1. Nonparametric EPSE Estimates for a Homogeneous Technology

Technology Type	Median Point Est.	*Categories of RS, %* DRS	CRS	IRS
Technology 1	0.987	1.6	20.0	78.3
Technology 2	0.995	8.4	28.3	63.2
Technology 3	0.994	4.8	34.2	61.0
Whole Sample	0.993	6.1	26.9	66.9

NOTE: Percentage points may not sum up to a hundred due to rounding.

output vector, evaluated in the [95%, 105%] interval of its output quantities. The definition of EPSE is such that values less/greater than unity are indicative of increasing/decreasing returns to scale. We note that EPSE, which relies on the estimates of the conditional mean as opposed to its gradients, is quite different from the elasticity-based measure of returns to scale in (4.1), based on which we draw our main conclusions in this paper.

Like before, we group credit unions into three returns to scale categories. A credit union is said to exhibit DRS/CRS/IRS if its EPSE point estimate is found to be statistically greater than/equal to/less than unity at the 95% significance level. We use wild bootstrap with 399 iterations to construct standard errors for the EPSE estimates. The break-down of the results by technology type is provided in the right panel of Table B.1. In a stark contrast to our preferred model 3, the EPSE estimates from a nonparametric specification indicate that a non-negligible fraction of credit unions in our sample (33.1%) exhibit non-IRS with the fraction growing substantially as one moves from Type 1 to Type 3 credit unions.[29] Rather surprisingly, the latter findings dramatically differ from those reported in Wheelock and Wilson (2011), who find no EPSE-based evidence against pervasive IRS amongst credit unions. Given that our nonparametric specification closely follows Wheelock and Wilson (2011), we suppose the primary reason why our EPSE estimates may still differ from theirs is because they (i) do not include financial capital as a quasi-fixed input and (ii) transform the raw data via the principal component extraction technique. Such a data transformation might lead to the loss of nonlinear information in the regressors which, in turn, might affect the estimates of scale economies. Further, the sensitivity of the scale economies estimates to the inclusion of equity capital in the model has also been documented in the case of commercial banks (Hughes and Mester, 1998). While the analysis of factors underlying the differences between our EPSE estimates and those reported in Wheelock and Wilson (2011) is unarguably of great empirical interest, their thorough investigation is well beyond the scope of the present paper.

References

Arellano, M., Bover, O., and Labeaga, J. M. (1999). Autoregressive models with sample selectivity for panel data. In Hsiao, C., Lahiri, K., Lee, L., and Pesaran, M., editors, *Analysis of Panels and Limited Dependent Variable Models*. Cambridge University Press, Cambridge.

Arellano, M. and Carrasco, R. (2003). Binary choice panel data models with predetermined variables. *Journal of Econometrics*, 115:125–157.

Baltagi, B. (2013). *Econometric Analysis of Panel Data*. Wiley, John & Sons Inc, 5th edition.

[29]The correspondence in rankings of credit unions across Wheelock and Wilson's (2011) model and ours is also weak, with the rank correlation coefficient of the returns to scale estimates from the two models being between 0.40 and 0.59.

Bauer, K. (2008). Detecting abnormal credit union performance. *Journal of Banking & Finance*, 32(4):573–586.

Bauer, K. J., Miles, L. L., and Nishikawa, T. (2009). The effect of mergers on credit union performance. *Journal of Banking & Finance*, 33(12):2267–2274.

Caves, D. W., Christensen, L. R., and Swanson, J. A. (1981). Productivity growth, scale economies, and capacity utilization in US railroads, 1955-74. *American Economic Review*, 71(5):994–1002.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47(1):225–238.

Charlier, E., Melenberg, B., and van Soest, A. (2001). An analysis of housing expenditure using semiparametric models and panel data. *Journal of Econometrics*, 101(1):71–107.

Dustmann, C. and Rochina-Barrachina, M. E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *Econometrics Journal*, 10(2):263–293.

Emmons, W. R. and Schmid, F. A. (1999). Credit unions and the common bond. *Federal Reserve Bank of St. Louis Review*, 81(September/October).

Frame, S. W. and Coelli, T. J. (2001). US financial services consolidation: The case of corporate credit unions. *Review of Industrial Organization*, 18(2):229–241.

Frame, S. W., Karels, G. V., and McClatchey, C. A. (2003). Do credit unions use their tax advantage to benefit members? Evidence from a cost function. *Review of Financial Economics*, 12(1):35–47.

Fried, H. O., Lovell, C., and Yaisawarng, S. (1999). The impact of mergers on credit union service provision. *Journal of Banking & Finance*, 23(2):367–386.

Gayle, G.-L. and Viauroux, C. (2007). Root-N consistent semiparametric estimators of a dynamic panel-sample-selection model. *Journal of Econometrics*, 141:179–212.

Goddard, J. A., McKillop, D. G., and Wilson, J. O. (2002). The growth of US credit unions. *Journal of Banking & Finance*, 26(12):2327–2356.

Goddard, J. A., McKillop, D. G., and Wilson, J. O. (2008). The diversification and financial performance of US credit unions. *Journal of Banking & Finance*, 32(9):1836–1849.

Hughes, J. P. and Mester, L. J. (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *Review of Economics and Statistics*, 80(2):314–325.

Hughes, J. P. and Mester, L. J. (2013). Who said large banks dont experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, 22(4):559–585.

Hughes, J. P. and Mester, L. J. (2014). Measuring the performance of banks: Theory, practice, evidence, and some policy implications. In Berger, A., Molyneux, P., and Wilson, J., editors, *Oxford Handbook of Banking*. Oxford University Press, Oxford, 2 edition.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65(6):1335–1364.

Lee, M.-j. and Vella, F. (2006). A semi-parametric estimator for censored selection models with endogeneity. *Journal of Econometrics*, 130:235–252.

Magnac, T. (2000). Subsidised training and youth employment: Distinguishing unobserved heterogeneity from state dependence in labour market histories. *Economic Journal*, 110(466):805–837.

Magnac, T. (2004). Panel binary variables and sufficiency: Generalized conditional logit. *Econometrica*, 72:1859–1876.

Malikov, E. and Kumbhakar, S. C. (2014). A generalized panel data switching regression model. *Economics Letters*, 124(3):353–357.

Malikov, E., Kumbhakar, S. C., and Sun, Y. (2016). Varying coefficient panel data model in the presence of endogenous selectivity and fixed effects. *Journal of Econometrics*, 190:233–251.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46:69–85.

National Credit Union Administration (2011). *2011 Annual Report*.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2):201–206.

Petersen, M. A. and Rajan, R. G. (2002). Does distance still matter? The information revolution in small business lending. *Journal of Finance*, 57(6):2533–2570.

Ridder, G. (1990). Attrition in multi-wave panel data. In Hartog, J., Ridder, G., and Theeuwes, J., editors, *Panel Data and Labor Market Studies*. North-Holland Publishing, Amsterdam.

Rochina-Barrachina, M. E. (1999). A new estimator for panel data sample selection models. *Annales d'Economie et de Statistique*, 55-56:153–181.

Semykina, A. and Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2):375–380.

Smith, D. J. (1984). A theoretic framework for the analysis of credit union decision making. *Journal of Finance*, 39(4):1155–1168.

Verbeek, M. (1990). On the estimation of a fixed effects model with selectivity bias. *Economics Letters*, 34(3):267–270.

Verbeek, M. and Nijman, T. E. (1996). Incomplete panels and selection bias. In Mátyás, L. and Sevestre, P., editors, *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, chapter 8. Kluwer Academic Publishers, Dordrecht.

Walter, J. R. (2006). Not your father's credit union. *Federal Reserve Bank of Richmond Economic Quarterly*, 92(4):353–377.

Wheelock, D. C. and Wilson, P. W. (2011). Are credit unions too small? *Review of Economics and Statistics*, 93(4):1343–1359.

Wheelock, D. C. and Wilson, P. W. (2013). The evolution of cost-productivity and efficiency among US credit unions. *Journal of Banking & Finance*, 37:75–88.

Wilcox, J. A. (2005). Economies of scale and continuing consolidation of credit unions. *FRBSF Economic Letter*, (Nov 4).

Wilcox, J. A. (2006). Performance divergence of large and small credit unions. *FRBSF Economic Letter*, (Aug 4).

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1):115–132.