Proxy Variable Estimation of Multi-Product Production Functions*

Emir Malikov[1] Gudbrand Lien[2]

[1]University of Nevada, Las Vegas
[2]Inland Norway University

February 10, 2021

Abstract

Productivity studies that use proxy variable estimators routinely specify single-output production functions despite most firms producing multiple outputs. This is usually accomplished by aggregating the firm's outputs using total revenue. Such a formulation rarely provides an adequate representation of the multi-product firm's production: (i) it cannot identify technological cross-output trade-offs along the production possibilities frontier, (ii) it unrealistically assumes perfect output substitutability which, if misspecified, produces biased estimates. We develop a new methodology for estimating multi-product production functions that addresses these limitations without requiring information beyond that which is available in most datasets and showcase it on Norwegian dairy-and-beef farming.

Keywords: multiple outputs, production function, proxy variable, structural identification, endogeneity, sieve estimation, B-splines, dairy and beef

JEL Classification: C14, D24, L10, Q12

**Correspondence to* Emir Malikov at emir.malikov@unlv.edu. We would like to thank the co-editor, Timothy Richards, and four anonymous referees for many helpful comments and suggestions that greatly helped improve the paper. We are also thankful to Henry Kinnucan, Subal Kumbhakar, Shunan Zhao, Mike Tsionas, seminar participants at Auburn University as well as conference participants at the 2016 AAEA Annual Meeting in Boston for feedback on earlier drafts.

1 Introduction

Applied economists routinely estimate production functions to understand technological relationships between inputs and outputs, to measure productivity growth or to quantify the effects of research and development, innovation, extension, etc. on firm productivity. This is critical for designing effective economic policy as well as for operations management at the firm level. However, identifying production functions from observational data is not a trivial task due to the endogeneity-inducing presence of latent firm productivity (think, managerial quality), and the literature has never ceased improving its measurement methods.[1] Still, many fundamental methodological issues remain outstanding, among which is estimating production functions for multi-product firms that are more the rule than the exception, especially in agricultural production. In this paper, we develop a new methodology for the proxy variable estimation of multi-product production functions which successfully identifies cross-output technological relationships within firms without requiring the access to information beyond that which is available in most datasets.

In light of the unsatisfactory performance of conventional approaches to tackling endogeneity in the production function context, such as fixed effects estimation or instrumenting using prices (see Ackerberg et al., 2015; Gandhi et al., 2020, for excellent discussions), the recently developed proxy variable approach to structurally identifying production functions by Olley & Pakes (1996) and Levinsohn & Petrin (2003) has gained wide popularity among practitioners. As a testament thereto, these two papers alone have accumulated over 12,500 citations as of the time of writing. However, most empirical studies applying such proxy variable estimators routinely specify *single*-output production functions[2] despite that, in practice, the majority of firms in both the developed and developing countries produce *multiple* outputs. For instance, in India's manufacturing, 47% of firms are multi-product and account for 80% of aggregate output (Goldberg et al., 2010) with the corresponding statistics for the U.S. being 39% and 87%, respectively (Bernard et al., 2010). Multi-output producers also dominate agriculture. The average number of commodities produced by U.S. corn farms is about 3.1, with 87% of the farmers also growing soybeans (Foreman, 2014). Multiple-output production is also very common among rural family farmers in underdeveloped countries, which is mainly driven by their pursuit of revenue and risk diversification and achieving self-sufficiency (e.g., see Mensah & Brümmer, 2016). Thus, multi-product firms are ubiquitous.

Arguably, the primary reason for the vast popularity of single-output proxy variable methods in the literature is the systematic lack of data on input usage by products. Individual products are customarily aggregated into the firm's single "output" using information on total revenue. While one can always employ such a single-product formulation of the production process, it rarely

[1] While unobserved by an econometrician, this latent productivity is one of the key determinants of the firm's endogenous input allocations resulting in its correlation with inputs inside the production function. If not accounted for explicitly, the presence of firm productivity therefore gives rise to the endogeneity-inducing omitted variable problem. Marschak & Andrews (1944) is among the first discussions of this simultaneity issue, which is also sometimes referred to as the "transmission bias" problem (see Griliches & Mairesse, 1998).

[2] For some influential applications and extensions of such proxy variable methods recently, e.g., see De Loecker & Warzynski (2012) or De Loecker et al. (2020).

provides an adequate representation of the multi-product firm's technology.

First, the total-revenue production function does not allow one to identify cross-output elasticities representing the technological trade-off between individual outputs along the firm's production possibilities frontier. Second and perhaps more importantly, such a single-output approach quite unrealistically assumes perfect substitutability of outputs effectively embedded in a linearly additive revenue-based output aggregator and arbitrarily sets aggregation weights equal to relative prices. In fact, if a revenue-based aggregator diverges from the unknown true aggregator in more than a constant (same for all firms), then a single-output model would likely produce biased estimates of both the firm's production technology and productivity. Third, formally justifying the use of total revenue as an output in the single-product production function for multi-product firms usually requires the strong assumptions of production technologies being identical and inputs being proportionally allocated across different product lines (see De Loecker, 2011).

With the above limitations of the total-revenue production function estimators in mind, we provide a new methodology for the proxy variable identification of production functions in a multi-product setting. Specifically, we consider the estimation of (single-equation) multi-product production functions where, motivated by Fernández et al. (2000), we do not aggregate outputs *a priori* but instead let them form an unspecified aggregator index function that represents a set of maximal quantities of all outputs producible from inputs given the firm's productivity. We tackle the endogenous static inputs in the presence of unobserved firm-level productivity by building upon Gandhi et al. (2020) whose system approach we extend to the case of multiple outputs. To make the estimation robust to the production technology misspecification, we employ a fully nonparametric formulation, which we operationalize using cubic B-spline sieves.

We show that in the case of multi-product firms the conventional assumptions for proxy variable estimators are insufficient to identify both the production function and production possibilities frontier, and some additional structural regularization is necessary. To structurally identify the multi-product model, we assume that (i) inputs exhibit some degree of product specificity and are not perfectly substitutable across product lines, with the implication being that (ii) before choosing inputs, the firm determines the degree of its relative specialization in each output by choosing the output ratios it intends to produce and these decisions are dynamic and costly, and (iii) the factor adjustment costs are subject to unanticipated shocks realized at the time of dynamic input choices.

The first two assumptions are useful because they imply that the firm predetermines not only the quantities of dynamic inputs but also their cross-product allocation, which we model as the firm's choice of output ratios. These relative output specialization decisions are also dynamic and *pre*cede those about input allocations. The latter renders the output *ratios* weakly exogenous and thus self-instrumentable. As a result, the output index function is immune to Gandhi et al.'s (2020) under-identification critique whereby the production model lacks valid excluded instruments for the endogenous regressors in violation of the "rank condition." The third assumption, while not being directly used in the estimation, guarantees that output ratios are not perfectly functionally dependent [in the spirit of Ackerberg et al.'s (2015) arguments] on dynamic inputs, and hence the output index function is separably identified.

Ours is not the only methodology for estimating the production technology and productivity of multi-product firms. Some recently developed estimators have attempted to extend the proxy variable identification to the multiple-output case by pursuing the estimation of heterogeneous within-firm product-specific production functions. However, they either are very data-intensive, requiring access to information on both single- and multiple-product firms (see De Loecker et al., 2016) or necessitate parameterizing product-specific demand functions in order to recover unobserved input allocations from the firm-year variation in output prices (see Valmari, 2016). In contrast to our proposed methodology, these (multiple-equation) production function models also cannot identify technological cross-output trade-offs, which are of great interest on their own.

The main advantages of our methodological contribution are therefore as follows. First, by formulating a single-equation representation of the multi-product production, our model is able to identify the technological relationship between individual outputs along the firm's production possibilities frontier, which both the (single-equation) total-revenue and the multiple-equation multi-product models are unable to deliver. Second, our methodology does not require information beyond that which is available in most production datasets so long as the product-specific output quantities and total firm-level inputs are observed. Lastly, as we explain in detail in the next section, the methodology that we develop also offers important benefits over other existing single-equation multi-product production function estimators, notably by Dhyne et al. (2016) and Grieco & McDevitt (2017), whose papers relate to ours most closely.

We showcase our methodology by applying it to recent data on Norwegian farms producing two products: milk and beef. Our choice of data is chiefly motivated by the availability of all relevant production variables, including the quantity—not just revenue—measures of outputs. Focusing on dairy farming for which beef is a (secondary) byproduct also aligns well with the underlying structural assumptions in our model. To empirically assess the sensitivity of the production function estimates to the treatment of outputs in multi-product processes, we also estimate the conventional single-output production function where the two outputs are *a priori* aggregated into a single revenue-based measure. The differences across the two models are non-negligible, with the potential of leading to dissimilar conclusions including those about scale efficiency and technological substitutability of products.

The rest of the paper proceeds as follows. We briefly discuss the related productivity literature in Section 2. Section 3 describes the multi-product production model along with the structural conceptualization of the firm's production decisions. We discuss the identification and estimation strategy in Section 4. Section 5 describes the data. The empirical application is discussed in Section 6 followed by concluding remarks in Section 7.

2 Background

Our methodology for the proxy variable identification of multi-product production functions contributes to the new but growing literature on the structural identification of multiple-output technologies. Like other proxy variable estimators, such papers rely on the use of lagged inputs/outputs as a source of exogenous variation (under some assumptions about the timing of firm's production decisions and its economic environment) to identify production functions. Among them, Dhyne et al. (2016) and Grieco & McDevitt (2017) are the two papers that relate to ours most closely. Both put forward production function estimators for multi-product firms that seemingly can also identify cross-output technological trade-offs without requiring data on prices or input allocations by product.

Like our paper, Grieco & McDevitt (2017) pursue a single-equation approach by making use of the output index function to formalize the multi-product inputs-to-outputs transformation. Dhyne et al. (2016) however opt for a multi-equation system of "generalized" production functions measuring the maximal quantity of a given output producible from the firm's total input quantities while holding quantities of other outputs fixed. Both estimators assume the Cobb-Douglas specification for the production function and thus, unlike ours, are not fully nonparametric. They also employ a proxy other than intermediate inputs to control for productivity (hiring or investment).

Our methodology offers some important benefits over these existing alternatives. Specifically, among other things, Dhyne et al.'s (2016) model may be susceptible to under-identification due to the lack of relevant excluded instruments for the outputs appearing on the right-hand side of their model (regardless of the chosen productivity proxy) as well as likely violation of the scalar unobservability condition necessary for invertability of the proxy. On the other hand, while being perfectly sound in their well-crafted setup, Grieco & McDevitt's (2017) estimator likely becomes no longer identified when using a freely varying input to proxy for latent productivity. This is especially unfortunate given that materials are the predominant proxy choice for production function estimators in applied literature owing to the impracticality of investment or hiring proxies because they tend to take too many zero values, rendering large chunks of data unusable.[3]

Furthermore, both Dhyne et al.'s (2016) and Grieco & McDevitt's (2017) models make it difficult to measure returns to scale because they have outputs appear on the right-hand side of the production function *in levels*. Hence, the estimated input gradients of the production function in their models measure changes in the left-hand side output while holding all other outputs constant. Thus, the sum of input elasticities in these models may provide a misleading measure of returns to "scale" with the latter traditionally referring to the change in *all* outputs. In contrast, outputs enter the right-hand side of our model in the ratio form ensuring that the estimated input elasticities readily measure the effects of input expansion not only on the left-hand-side output but on all outputs proportionately, which is in line with the conventional multi-output definition of returns

[3]To conserve space, the detailed derivation of the stated limitations of Dhyne et al.'s (2016) and Grieco & McDevitt's (2017) estimators is omitted from the paper. It is available in the Online Supplementary Material.

to scale.

Beyond the papers on the *proxy variable* identification of production functions, the literature on primal estimation of multi-output production technologies is vast and well established. It is also quite heterogeneous in the methodological paradigms adopted by researchers. Methodologies range from the axiomatic/deterministic mathematical-programming-based “data envelopment analysis” (for excellent reviews, see Ray, 2004; Charnes et al., 2013) to the “stochastic frontier analysis” that allows for random shocks (e.g., Kumbhakar & Lovell, 2000; Kumbhakar, 2013).[4] Among these two strands of literature, identification is of concern mainly for stochastic frontier models which involve econometric (non-axiomatic) estimation.

The proxy variable identification strategy that we pursue in this paper conceptually differs from the approach to handling endogeneity that is popular in stochastic frontier literature and relies on firm-varying output/input prices as the source of exogenous variation (e.g., Atkinson & Primont, 2002; Atkinson et al., 2003; Tsionas et al., 2015; Kumbhakar & Tsionas, 2016; Atkinson & Tsionas, 2016). Exogenous price variation is used for identification both directly as instruments and indirectly when formulating the full- or limited-information maximum likelihood system of equations. In many instances, these stochastic frontier models also require additional assumptions such as the parameterization of the relationship between the endogenous regressors and the stochastic component in the model (e.g., Kutlu, 2010; Tran & Tsionas, 2013; Griffiths & Hajargasht, 2016).

The validity of using price information to identify the firm’s production technology, however, is not obvious. Not only are the price data often unavailable or prone to measurement errors (Levinsohn & Petrin, 2003), but the use of prices may also be problematic on theoretical grounds (see Griliches & Mairesse, 1998; Ackerberg et al., 2007, 2015). Specifically, the validity of prices as exogenous instruments is normally justified by invoking the assumption of perfectly competitive markets. However, if firms were indeed price-takers, in theory, one should not observe the firm-level variation in prices and, without such a variation, prices cannot be used as operational instruments. Even with the prices varying exogenously across space, such a variation will likely be insufficient to identify the nonparametric production function because the instrumental-variable identification in nonparametric models generally requires an instrument to be as complex as the endogenous variable it is meant to instrument (Newey & Powell, 2003). If a researcher does observe variation in prices across individual firms, the latter variation likely reflects differences in their market power and/or the quality of inputs/outputs. In both cases, the variation in prices is unlikely to be exogenous to the firm’s decisions and hence cannot help the identification (see Gandhi et al., 2020). In this paper, we therefore pursue an alternative identification strategy which does not require *firm*-level variation in prices.[5]

It is also noteworthy that the identification strategies pursued in the stochastic frontier literature

[4] Although, the deterministic-vs-stochastic dichotomy between the two methodological paradigms is becoming increasingly moot in light of the recently developed stochastic generalizations of the axiomatic data envelopment analysis (e.g., see Kuosmanen et al., 2015).

[5] Few available methods in the stochastic frontier literature that do not require the price information for identification make use of either higher moment conditions or copulas (e.g., see Tran & Tsionas, 2015).

are developed almost exclusively under the assumption that the endogeneity of regressors arises from their correlation with the random shock as opposed to latent firm efficiency/productivity, thereby rather unrealistically maintaining that firm productivity is random and unknown not only to an econometrician but also to the firm. Few notable exceptions include Amsler et al. (2016, 2017) but they too require availability of external instruments for identification.

3 Multi-Product Production

Consider the production process of a multi-product firm i ($i = 1, \dots, n$) in the time period t ($t = 1, \dots, T$) in which physical capital K_{it}, land N_{it}, labor L_{it} and the intermediate input M_{it} are transformed into $S \geq 1$ outputs $(Y_{1,it}, \dots, Y_{S,it})$ via the production function $F(\cdot)$ given the firm's productivity. To model this multi-product production, we generalize the single/total-output formulation of the firm's production process customarily employed in the literature to the case of multiple outputs as follows:

$$H(Y_{1,it}, \dots, Y_{S,it}) = F(K_{it}, N_{it}, L_{it}, M_{it}) \exp\{\omega_{it} + \eta_{it}\}, \tag{3.1}$$

where, following Fernández et al. (2000) and Fernández et al. (2002, 2005),[6] the outputs are combined into an aggregate output index function $H(\cdot)$ which essentially represents the firm's production possibilities frontier, i.e., the combinations of maximal quantities of all outputs producible from inputs given the level of firm productivity. As usual, $(\omega_{it} + \eta_{it})$ is the firm's composite productivity term consisting of the persistent productivity ω_{it} and a random transitory shock η_{it}. Both inputs and productivity are defined at the firm level. In what follows, we characterize structural (conceptual) assumptions about the firm's technology, economic environment and its decision-making process which facilitate the structural identification of the model.

Assumption 1 *The production relationship between inputs and outputs takes the form of* (3.1). *(i) The output index $H(\cdot)$ is multiplicatively separable from the production function $F(\cdot)$ and differentiable, positively monotone and linearly homogeneous in outputs. (ii) The production function $F(\cdot)$ is continuous and satisfies the standard neo-classical assumptions, including differentiability, positive monotonicity and concavity in inputs. (iii) Persistent and transitory productivity ω_{it} and η_{it} are Hicks-neutral, log-additive and defined at the firm level.*

While at first the separability of functions $H(\cdot)$ and $F(\cdot)$ may seem to be rather restrictive, it is also imposed onto a standard single-output production function where the (single) total output is assumed to be separable from $F(\cdot)$.[7] Clearly, (3.1) nests the single-product formulation as a special case, i.e., the case when $S = 1$ and function $H(\cdot)$ is normalized to an identity link function.

[6]Unlike them, however, we do not parameterize neither function $F_t(\cdot)$ nor function $H_t(\cdot)$ leaving them to be unspecified nonparametric functions.

[7]Note that in some applications, where different outputs require a different input mix, the separability may be quite unappealing. Should that be the case, the use of the proposed model is left up to a researcher's discretion.

Some recent specifications of similar separable multi-product production functions include Greene (2008), O'Donnell & Nguyen (2013) and O'Donnell (2014) among others, although none of these studies focus on structural identification.

To ensure a meaningful representation of the firm's production technology, we also assume that the output index $H(\cdot)$ is positively monotone in outputs. The monotonicity assumption is rather natural and ensures that outputs are increasing in inputs and productivity. It also implies that outputs are net substitutes, which effectively rules out the case of the so-called "bad" outputs, such as pollution, that tend to be complementary to the production of intended "good" outputs. The assumption about homogeneity of $H(\cdot)$ is essentially a normalizing restriction required to identify the output index function which ensures the latter is available for use as a sufficient statistic for the unknown production function $F(\cdot)$. This is analogous to normalizing to unity the unknown coefficient in front of the total output appearing on the left-hand side of the single-output production function. The benefit of such a normalizing restriction over, say, the alternative assumption of multiplicative separability of $H(\cdot)$ in outputs implicitly employed by Grieco & McDevitt (2017) is the invariance of output index to the choice of a normalizing output.

The assumption that firm productivity is Hicks-neutral is standard in productivity literature and dates back to Olley & Pakes (1996). We also assume that both the persistent productivity ω_{it} and the transitory productivity shock η_{it} are firm-specific, implying that the multi-product firm produces all its S outputs with the same level of productivity. In doing so, we follow popular practice in modeling multi-product firms (e.g., De Loecker, 2011; De Loecker et al., 2016). This assumption seems appropriate especially given that ω_{it} is normally interpreted as capturing unobserved firm characteristics such as management quality and tacit knowledge that are important for the production of all outputs.

Assumption 2 *Firms maximize the discounted stream of life-time profits in perfectly competitive output and intermediate input markets.*

Following the bulk of the literature, we assume perfectly competitive markets implying that firms are price-takers which, at least in theory, rules out any operational firm-level variation in prices. In what follows, we therefore omit prices from the list of relevant determinants entering the firm's decision equations. As discussed in Introduction, this paper seeks to develop an estimator that does not require firm-level price information.

Assumption 3 *Among the firm's inputs: (i) physical capital K_{it}, land N_{it} and labor L_{it} are dynamic inputs subject to adjustment frictions; (ii) intermediate inputs M_{it} are freely varying, and the corresponding demand function is strictly monotone in a scalar unobservable ω_{it}; (iii) all inputs exhibit some degree of product specificity.*

Similar to Gandhi et al. (2020) and Malikov et al. (2020), we assume that, unlike freely varying intermediate inputs M_{it}, the capital K_{it} and labor L_{it} are subject to adjustment frictions (e.g., time-to-install, hiring costs) and are quasi-fixed.[8] It is perhaps even more natural to assume that land N_{it} is quasi-fixed too. Inputs K_{it}, N_{it} and L_{it} are the state variables with dynamic implications. Intermediate inputs M_{it} are freely varying and are chosen by the firm statically, i.e., with no dynamic implications. Along the lines of Levinsohn & Petrin (2003), Ackerberg et al. (2015) and Gandhi et al. (2020), we also assume strict monotonicity of the firm's demand function for M_{it} in a scalar unobservable ω_{it}.

By Assumption 3, all inputs are also said to be product-specific to a degree, and hence they are not perfectly substitutable across product lines. Thus, their use within the firm depends on the product designation. This assumption is not unreasonable given that the production lines for different outputs within the firm are likely to be separable to a certain degree, implying that inputs cannot be freely moved across product lines without incurring some reallocation costs (e.g., recalibration of the machinery, retraining of the personnel). This implies that *before* the firm chooses input quantities it first needs to choose the degree of its relative specialization in each output that would then dictate the allocation of inputs across product lines. We model these relative specialization decisions (and implicitly, the cross-product allocations) as the firm's choice of its output mix captured by the *output ratios* it intends to produce. Given the product-specific nature of inputs and the costly cross-product reallocation thereof, the choice of the firm's output mix, as captured by its output ratios, is a delayed dynamic decision rendering the output ratios quasi-fixed variables just like the dynamic inputs. Kumbhakar (2013) makes a similar argument, although in a somewhat different context, whereby output ratios in a separable radial multi-output distance function are deemed quasi-fixed and thus weakly exogenous under the (albeit static) profit maximization premise. We summarize this explicitly in the following assumption.

Assumption 4 *Before choosing the input quantities, the firm determines the degree of its relative specialization in each output by choosing the output ratios it intends to produce. These decisions are dynamic and subject to adjustment costs.*

As shown in Section 4, without explicitly conceptualizing how the firm determines its output mix configuration, one generally cannot identify the production technology of multi-product firms under the standard assumptions. For instance, recognizing the latter, Grieco & McDevitt (2017) adopt a model whereby firms predetermine their output quantities (i.e., levels) in advance. The conceptual framework upon which we rely in this paper is less restrictive, however, as unlike Grieco & McDevitt (2017), we assume that the firm chooses not the output *levels* but rather output *ratios* thereby letting output quantities change in the face of updates in persistent productivity via freely varying input decisions as well as random productivity shocks in the next period when these outputs are actually produced.

Together, Assumptions 3–4 formalize not only the timing of production decisions by the firm but also their relative order. For concreteness, we split the evolution between time periods $t-1$

[8] This timing assumption about L_{it} seems reasonable given the microeconomic evidence in support of quasi-fixity of labor in the wake of nonlinear adjustment costs (e.g., Caballero et al., 1997; Cooper & Willis, 2003).

and t into two subperiods: from $t-1$ to $t-b$ and from $t-b$ to t, where $b \in (0,1)$. In this timeline, (i) the decision about relative specialization in period t is made at time $t-1$, followed by the choice of dynamic inputs for period t at time $t-b$; and (ii) the choice of freely varying inputs for the period t production is contemporaneous. In $b \neq 0$, we have that the dynamic input choices are predetermined just like those of the output ratios; in $b \neq 1$, we have that the output ratios are chosen by the firm before it optimizes all inputs.

Because this order/timing assumption is untestable, it is imperative that researchers critically assess its suitability for their empirical application. Namely, the assumption about the firm's choice of the relative output specialization preceding its input choices requires that, while product-specific to a degree, the inputs be yet not *completely* specialized by product whereby output-specific production decisions could be made independently. For example, in a manufacturing factory with multiple product lines this assumption may not be reasonable because individual production lines can usually be sped up or slowed down independently. On the other hand, in an oil refinery where different fuels are produced as byproducts of the same refining process, our assumption is more likely to hold. In the case of our own empirical application, this assumption is plausible because beef and dairy products are natural byproducts. Other agricultural applications such as crop production also fit our model setup, whereby farmers' crop mix choices inform how they then allocate land use to various crops, followed accordingly by the purchasing decisions for pesticides and fertilizers (intermediate inputs) during the growing season.

However, the timing and order assumption about production decisions alone is insufficient to structurally identify multi-product technology. Because the output ratio decision is in principle no different from the firm's choice of dynamic inputs, in order to distinguish between the two, it is imperative that they contain independent variation. This is important for ruling out "perfect functional dependence" between the output ratios and the inputs in order to identify the model (for a detailed discussion, see Section 4). Therefore, we supplement Assumption 4 with an additional structural assumption about differential information sets available to the firm during decision-making. We formalize the latter in terms of the firm's knowledge of transitory shocks to the factor adjustment costs it must incur.

Namely, we assume that besides the within-firm considerations the adjustment costs for dynamic inputs (capital, land and labor) are also influenced by a firm-specific exogenous "shock," say ξ, representative of external factors that are relevant for capital and land investments and hiring decisions. This shock is *i.i.d.* over time: unanticipated until its realization at the time of dynamic input adjustments. Thus, in the timeline of our model, ξ_{it-b} is observed by the firm at time $t-b$ after the choice of relative output specialization for period t had already been made at time $t-1$. Examples of such a cost shock include the "access-to-finance" shock, availability of workers, utilization shocks, local institutional shocks, extreme weather shocks, etc. As we show in Section 4, the presence of an exogenous random shock ξ_{it-b} that generates independent variation lets us circumvent the Ackerberg et al. (2015) critique and successfully identify the multi-product output index function. Crucially, ξ_{it-b} need *not* be observed for the estimation purposes.

Relying on additional structural assumptions, including those about differential information sets, to break perfect functional dependence between variables in the production function is not unprecedented and has been first suggested by Ackerberg et al. (2015). For instance, they consider a possibility of the within-period evolution of persistent productivity with an intermediate update thereto at the similarly defined time $t-b$. Grieco & McDevitt (2017) also resort to an analogous assumption about the within-period evolution of ω_{it} along with the timing of the firm's output decisions to ensure some independent variation in the target levels of outputs. In this paper, we take a somewhat different route, whereby we continue to maintain the usual assumption about ω_{it} updating once in period t as it evolves from period $t-1$. The differential information sets are ensured by the presence of an unanticipated cost shock that provides an intermediate update to the firm's information set.

Let Ξ_{it} be the information set available to the firm for making the time t decisions. Then, the dynamics of unobservables are summarized as follows.

Assumption 5 *(i) Adjustment cost shock ξ_{it} is i.i.d. over t and $\mathcal{P}_\xi(\xi_{it}|\Xi_{it-a}) = \mathcal{P}_\xi(\xi_{it})$ for $a > 0$, with the realizations at the time of dynamic input adjustments. (ii) Persistent productivity ω_{it} evolves between two consecutive time periods according to the time-homogeneous exogenous first-order Markov process: $\mathcal{P}_\omega(\omega_{it}|\Xi_{it-a}) = \mathcal{P}_\omega(\omega_{it}|\omega_{it-1})$ for $0 < a \leq 1$. (iii) The transitory productivity shock η_{it} is i.i.d. over i and t, and $\mathcal{P}_\eta(\eta_{it}|\Xi_{it}) = \mathcal{P}_\eta(\eta_{it})$.*

We now formalize the firm's optimization problem. Without loss of generality, let the firm's output ratios be defined using $Y_{1,it}$ as a base: $Z_{s,it} \stackrel{\text{def}}{=} Y_{s,it}/Y_{1,it}\ \forall\ s = 2,\ldots,S$. Given the dynamic nature of the output ratios, they play a role of additional state variables besides dynamic inputs K_{it}, N_{it} and L_{it}. All dynamic variables evolve between time periods according to their respective deterministic controlled laws of motion. Although such laws of motion are normally postulated as the evolution process from period $t-1$ to period t, they can easily be decomposed into two subperiods in order to match the intra-period timing and order assumptions about the firm's choices of the state variables. Namely, (i) the evolution of dynamic variables from period $t-1$ to subperiod $t-b$ is given by $K_{it-b} = K_{it-1}$, $N_{it-b} = N_{it-1}$, $L_{it-b} = L_{it-1}$ and $Z_{s,it-b} = Z_{s,it-1} + \Delta Z_{s,it-1}$, and (ii) from subperiod $t-b$ to period t is given by $K_{it} = (1-\delta)K_{it-b} + I_{it-b} = (1-\delta)K_{it-1} + I_{it-b}$, $N_{it} = N_{it-b} + \Delta N_{it-b} = N_{it-1} + \Delta N_{it-b}$, $L_{it} = L_{it-b} + \Delta L_{it-b} = L_{it-1} + \Delta L_{it-b}$ and $Z_{s,it} = Z_{s,it-b}$, where δ is the depreciation rate, I_{it-b} is the gross investment in physical capital, ΔN_{it-b} is the net change in land holdings, and ΔL_{it-b} represents the net hiring.

For convenience, let $\mathbf{Z}_{it} = (Z_{2,it},\ldots,Z_{S,it})'$ with $\Delta\mathbf{Z}_{it} = (\Delta Z_{2,it},\ldots,\Delta Z_{S,it})$ for all t. Analogously, let $\mathbf{X}_{it} = (K_{it}, N_{it}, L_{it})'$ for all t and $\Delta\mathbf{X}_{it-b} = (I_{it-b}, \Delta N_{it-b}, \Delta L_{it-b})'$ in all intermediate subperiods $t-b$. The firm's dynamic optimization problem is then described by the following Bellman equation for the choice of output ratios $\mathbf{Z}_{it}$:

$$\mathbb{V}_{t-1}\big(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1}\big) = \sup_{\Delta\mathbf{Z}_{it-1}} \Bigg\{ \Pi_{t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1}) - \mathrm{C}^{\mathcal{Z}}_{t-1}(\Delta\mathbf{Z}_{it-1}, \mathbf{Z}_{it-1}) +$$

$$\mathbb{E}\left[\mathbb{V}_{t-b}\big(\mathbf{X}_{it-1}, \underbrace{\mathbf{Z}_{it-1}+\Delta\mathbf{Z}_{it-1}}_{\mathbf{Z}_{it-b}}, \omega_{it-1}\big)\Big|\ \Xi_{it-1}\right]\Bigg\}, \tag{3.2}$$

along with the Bellman equation for the choice of dynamic inputs $\mathbf{X}_{it}$:

$$\begin{aligned}\mathbb{V}_{t-b}\big(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \omega_{it-1}\big) = \sup_{\Delta\mathbf{X}_{it-b}} \Bigg\{ & -\mathrm{C}^{\mathcal{I}}_{t-b}(I_{it-b}, K_{it-1}, \xi_{it-b}) - \mathrm{C}^{\mathcal{N}}_{t-b}(\Delta N_{it-b}, N_{it-1}, \xi_{it-b}) \\ -\mathrm{C}^{\mathcal{L}}_{t-b}(\Delta L_{it-b}, L_{it-1}, \xi_{it-b}) + \beta\mathbb{E}&\left[\mathbb{V}_t\Big(\underbrace{(1-\delta)K_{it-1}+I_{it-b}}_{K_{it}}, \underbrace{N_{it-1}+\Delta N_{it-b}}_{N_{it}}, \underbrace{L_{it-1}+\Delta L_{it-b}}_{L_{it}}, \mathbf{Z}_{it}, \omega_{it}\Big)\Big|\ \Xi_{it-b}\right]\Bigg\},\end{aligned} \tag{3.3}$$

where $\Pi_{t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1})$ is the restricted expected profit function with the freely varying intermediate inputs M_{it-1} being the choice variable in which it is maximized conditional on the already optimized quasi-fixed $(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1})$ and the observed level of productivity ω_{it-1}; and $\mathrm{C}^{\kappa}_{\tau}(\cdot)$ is the adjustment cost function for output specialization ($\kappa = \mathcal{Z}$), capital ($\kappa = \mathcal{I}$), land ($\kappa = \mathcal{N}$) and labor ($\kappa = \mathcal{L}$) at time τ.

Solving (3.2)–(3.3) for controls yields the optimal policy functions: the gross investment function $I_{it-b} = \mathbb{I}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-1})$, the net land demand $\Delta N_{it-b} = \mathbb{N}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-b})$ and the net hiring function $\Delta L_{it-b} = \mathbb{L}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-b})$ that all depend on $\xi_{it-b} \in \Xi_{it-b}$, and the relative specialization adjustment functions $\Delta Z_{s,it-1} = \mathbb{Z}_{s,t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1})\ \forall s$ that are unaffected by ξ_{it-b} because $\xi_{it-b} \notin \Xi_{it-1}$. Analogously, the solution to the restricted expected profit maximization corresponding to the value function $\Pi_{t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1})$ yields the conditional demand for intermediate inputs: $M_{it-1} = \mathbb{M}_{t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1})$. We provide an explicit form of this static optimization problem in Section 4.

Lastly, under Assumption 5(ii), we can express persistent productivity as

$$\omega_{it} = \psi(\omega_{it-1}) + \zeta_{it}, \quad \text{with } \mathbb{E}[\zeta_{it}|\Xi_{it-a}] = 0 \text{ for } a > 0, \tag{3.4}$$

where $\psi(\cdot)$ is an unspecified conditional mean function, and ζ_{it} is the unanticipated innovation in the firm's persistent productivity at time t, and note that $\mathbb{E}[\zeta_{it}|\Xi_{it-1}] = \mathbb{E}[\zeta_{it}|\Xi_{it-b}] = 0$. Also, under Assumption 5(iii), we normalize

$$\mathbb{E}[\eta_{it}|\Xi_{it}] = \mathbb{E}[\eta_{it}] = 0. \tag{3.5}$$

Remark 1 In Assumption 5 and the Bellman equations in (3.2)–(3.3), we assume that the firm is subject to a scalar shock that affects adjustment costs for all dynamic inputs. One can replace such a setup with a scenario where each dynamic input is subject to its own input-specific adjustment cost shock or, alternatively, the scenario where only some of the three dynamic inputs are subject to such a shock. Regardless of these specifics, the optimal policy functions for *all* dynamic inputs will generally continue to depend on the adjustment cost shock(s) because the firm learns this new information at the time when all these decisions are being made. Thus, the desired outcome would still be accomplished: all three I_{it-b}, ΔN_{it-b} and ΔL_{it-b} (and hence, K_{it}, N_{it} and L_{it}) will contain independent variation absent from $\{\Delta \mathbf{Z}_{s,it-1}\}$ (and thus from $\{\mathbf{Z}_{s,it}\}$).

Remark 2 Unlike a single-output model which uses the total revenue to measure the multi-product firm's outputs, the multi-output model in (3.1) does *not* unrealistically assume perfect substitutability of outputs effectively embedded in a linearly additive revenue-based output aggregator and does not arbitrarily set aggregation weights equal to relative prices. In light of Maasoumi & Nickelsburg's (1988) work on multidimensional "ideal" indices, if one were to *a priori* aggregate outputs, a CES aggregator (of which a linearly additive aggregator is a special case) would be a more preferable choice. Still yet, appropriate distributive weights and the substitution elasticity parameter for such an aggregator would need be estimated (also see Maasoumi & Racine, 2016). Instead, model (3.1) allows the output aggregator $H(\cdot)$ remain an unknown function to be identified from the data thereby reducing the risks of misspecifying the firm's technology. In fact, if the revenue-based aggregator $R(Y_{1,it},\dots,Y_{S,it}) \equiv \sum_s P_{s,t}Y_{s,it}$ diverges from unknown "true" aggregator $H(Y_{1,it},\dots,Y_{S,it})$ in more than a constant that is the same across all i and t, a single-output model would likely produce biased estimates of the firm's production function and productivity. More concretely, consider the case when $H(Y_{1,it},\dots,Y_{S,it}) \neq R(Y_{1,it},\dots,Y_{S,it})$ and define the difference between the two aggregators as $\Delta(\cdot) \equiv H(\cdot) - R(\cdot)$. Given the linear homogeneity of its two components, $\Delta(\cdot)$ is also linearly homogeneous. Then, the production function in (3.1) can be rewritten in logs as

$$\ln R(Y_{1,it},\dots,Y_{S,it}) = \ln F(K_{it},N_{it},L_{it},M_{it}) - \ln\left(1 + B(Y_{1,it},\dots,Y_{S,it})\right) + \omega_{it} + \eta_{it} \qquad (3.6)$$

where $B(\cdot) \stackrel{\text{def}}{=} \Delta(\cdot)/R(\cdot)$. From (3.6), it is evident that a revenue-based single-output model, which replaces $R(\cdot)$ on the left-hand side with observable revenue data, generally suffers from an omitted variable bias whereby the term $\ln\left(1 + B(Y_{1,it},\dots,Y_{S,it})\right)$ is absorbed in the error term mistakenly assumed to be *i.i.d.*. Since output quantities are correlated with both inputs and latent productivity, the estimates of $F(\cdot)$ and ω_{it} will then likely be biased.

4 Identification and Nonparametric Estimation

Estimating the production function in (3.1) is challenging not only due to the presence of multiple outputs but also because of the latent nature of ω_{it}. Omitting it from the regression is not an option, since the latter would give rise to the endogeneity problem given that ω_{it} is correlated with inputs. We tackle this problem by adopting a proxy variable approach à la Olley & Pakes (1996) and Levinsohn & Petrin (2003). Specifically, we consider the identification and estimation of the multi-product production function in (3.1) by building on Gandhi et al.'s (2020) methodology, which we generalize to accommodate multiple outputs.

Making use of the linear homogeneity of $H(\cdot)$ in outputs, for any $\kappa > 0$, we get

$$H(\kappa Y_{1,it}, \ldots, \kappa Y_{S,it}) = \kappa H(Y_{1,it}, \ldots, Y_{S,it}) = \kappa F(\mathbf{X}_{it}, M_{it}) \exp\{\omega_{it} + \eta_{it}\}, \tag{4.1}$$

where we have made a substitution for $H(\cdot)$ in the second equality using (3.1), and recall that $\mathbf{X}_{it} = (K_{it}, N_{it}, L_{it})'$ is a vector of quasi-fixed dynamic inputs. Setting κ equal to the *inverse* of one of the outputs, say $Y_{1,it}$, yields

$$Y_{1,it} = \frac{F(\mathbf{X}_{it}, M_{it})}{H\left(1, \frac{Y_{2,it}}{Y_{1,it}}, \ldots, \frac{Y_{S,it}}{Y_{1,it}}\right)} \exp\{\omega_{it} + \eta_{it}\}. \tag{4.2}$$

Taking logs of both sides of (4.2), we get

$$\begin{aligned} y_{1,it} &= f(\mathbf{X}_{it}, M_{it}) - h(1, Z_{2,it}, \ldots, Z_{S,it}) + \omega_{it} + \eta_{it} \\ &\stackrel{\text{def}}{=} f(\mathbf{X}_{it}, M_{it}) + \widetilde{h}(\mathbf{Z}_{it}) + \omega_{it} + \eta_{it}, \end{aligned} \tag{4.3}$$

where the lower-case variables/functions denote the logs of the respective variables/functions, the output ratios $\{Z_{s,it}\}$ are as defined earlier, and $\widetilde{h}(\cdot) \stackrel{\text{def}}{=} -h(1, \cdot)$.

4.1 Unidentification of the Standard Proxy Approach

Normally, to estimate (4.3) via proxy variable approach, one makes use of the Markovian nature of unobservable ω_{it} and then proxies for it by inverting one the firm's optimal policy functions. For instance, using (3.4), we can rewrite (4.3) as

$$\begin{aligned} y_{1,it} &= f(\mathbf{X}_{it}, M_{it}) + \widetilde{h}(\mathbf{Z}_{it}) + \psi(\omega_{it-1}) + \zeta_{it} + \eta_{it} \\ &\stackrel{\text{def}}{=} f(\mathbf{X}_{it}, M_{it}) + \widetilde{h}(\mathbf{Z}_{it}) + \psi\left(\mathbb{M}_{t-1}^{-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, M_{it-1})\right) + \zeta_{it} + \eta_{it}, \end{aligned} \tag{4.4}$$

where, assuming strict monotonicity of the firm's demand for intermediate inputs $\mathbb{M}_{t-1}(\cdot)|M_{it-1} > 0$ in a scalar unobservable ω_{it-1} for a given $(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1})$ per Assumption 3(ii), the function is inverted to proxy for ω_{it-1} in the last equality.

In what follows, we show that a proxy-based model in (4.4) generally is *not* identified thereby hindering its direct one-step consistent estimation. We will then proceed with the alternative identification approach.

Setting the issue of separable identifiability of the output index $\widetilde{h}(\cdot)$ aside [more on this later], the problem with (4.4) concerns unidentification of the production function $f(\cdot)$. Under the model assumptions, all right-hand-side covariates in (4.4) except for M_{it} are predetermined, implying their weak exogeneity with respect to both ζ_{it} and η_{it} which are unknown to the firm at time $t-a$ for $a > 0$, i.e., $\mathbb{E}\left[\zeta_{it} + \eta_{it} \,|\, \mathbf{X}_{it}, \mathbf{Z}_{it}, \mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, M_{it-1}\right] = 0$, and hence they can self-instrument. Consequently, identifying the production function $f(\cdot)$ in (4.4) requires the endogenous M_{it} be instrumented. Following the bulk of the literature, one may think that the function can seemingly be identified by utilizing abundant lagged inputs and possibly output ratios to instrument for M_{it} given that these lags are predetermined under assumptions about the evolution of firm productivity and hence also meet the weak exogeneity requirement for valid instruments. However, the identification can only be achieved if these additional lags provide some additional *relevant* (exogenous) variation for M_{it} after conditioning on the already included self-instrumenting variables. Intuitively, these additional lags must be relevant to meet the rank condition for identification of the model.

It happens that, despite the apparent abundance of instruments, the production function $f(\cdot)$ remains *un*identified in (4.4) because all relevant instruments for M_{it} that were initially excluded from (4.3) are now used to proxy for the unobserved ω_{it-1}. This is the key point of the Gandhi et al. (2020) critique of proxy estimators. More formally, the endogenous M_{it} entering $f(\cdot)$:

$$\begin{aligned} M_{it} &= \mathbb{M}_t(\mathbf{X}_{it}, \mathbf{Z}_{it}, \omega_{it}) = \mathbb{M}_t(\mathbf{X}_{it}, \mathbf{Z}_{it}, \psi(\omega_{it-1}) + \zeta_{it}) \\ &= \mathbb{M}_t\left(\mathbf{X}_{it}, \mathbf{Z}_{it}, \psi\left(\mathbb{M}_{t-1}^{-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, M_{it-1})\right) + \zeta_{it}\right), \end{aligned} \tag{4.5}$$

is a function of the following observables $(\mathbf{X}_{it}, \mathbf{Z}_{it}, \mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, M_{it-1})$ and the unobservable ζ_{it}. Comparing these arguments of M_{it} with the variables entering $f(\cdot)$ directly as well as appearing inside the proxy function $\psi(\cdot)$, it is evident that the only extra source of variation for M_{it}, which has not already been included on the right-hand side of (4.4), is the unobservable ζ_{it}. In other words, conditional on the already included self-instrumenting variables, there is *no* other relevant exogenous variable from within the production function that may be used to instrument for the endogenous M_{it} because, for it to be relevant in predicting M_{it}, it would have to correlate with ζ_{it}, which is the only source of "free" variation left in M_{it}. The correlation with ζ_{it} would however violate the exogeneity requirement thereby invalidating the instrument. Notably, this also applies to weakly exogenous excluded lags (of the second and higher order) of inputs and output ratios which, following the bulk of the literature, one may be tempted to utilize in hopes of identifying the model. Therefore, M_{it} lacks an excluded *relevant* instrument, and the production function $f(\cdot)$ in (4.4) is *un*identified.

To tackle these unidentification problems, we build upon Gandhi et al. (2020) whose approach we extend to the case of multiple outputs. Effectively, we dispense with the proxy model in (4.4) in favor of (4.3) which we augment with an equation for the optimality condition for intermediate inputs M_{it} derived from the firm's static profit-maximization problem and an equation for the Markov process of ω_{it}. Intuitively, this is akin to the system of simultaneous equations approach. While this solves the endogeneity of M_{it} aiding identification of the production function $f(\cdot)$, identifying the output index function $\widetilde{h}(\cdot)$ also requires that, besides being weakly exogenous, the output ratios $\mathbf{Z}_{it}$ be *not* "perfectly functionally dependent"[9] on the self-instrumenting dynamic inputs also entering the estimating equation, along the lines of the Ackerberg et al. (2015) critique. In what follows, we expand on this issue.

[9]In a linear parametric framework, "perfect functional dependence" is normally referred to as "perfect collinearity".

4.2 A System Approach to Identification

Since intermediate inputs are non-dynamic and freely varying, the firm's optimal choice of M_{it} can be modeled statically as the concentrated expected profit-maximization problem[10] subject to the (already predetermined) optimal choice of all dynamic inputs and output ratios, i.e.,

$$\max_{M_{it}} \sum_{j=1}^{S} P_{j,t}\mathbb{E}[Y_{j,it}|\ \Xi_{it}] - P_{M,t}M_{it} =$$
$$\max_{M_{it}} \sum_{j=1}^{S} P_{j,t}\frac{F(\mathbf{X}_{it}, M_{it})\exp\{\omega_{it}\}\mathbb{E}[\exp\{\eta_{it}\}|\ \Xi_{it}]}{\widetilde{H}^j(Z^j_{1,it}, \dots, Z^j_{S,it})} - P_{M,t}M_{it}, \tag{4.6}$$

where we have expressed each output $Y_{j,it}$ using the linear homogeneity property à la (4.2) under the respective $\kappa = 1/Y_{j,it}$ normalization. We add the superscript j to the normalized output index $\widetilde{H}(\cdot)$ as well as the output ratios $Z_{s,it}$ to explicitly recognize that the index and the ratios are defined differently each time, depending on the choice of normalizing output. Further, $P_{j,t}$ and $P_{M,t}$ are the prices of output j and materials, respectively; these prices do *not* need to vary across firms. The value function corresponding to (4.6) at time $t-1$ yields $\Pi_{t-1}(\cdot)$ entering the firm's Bellman equation (3.2).

The first-order condition with respect to M_{it} is given by

$$\sum_{j=1}^{S} P_{j,t}\frac{\partial F(\mathbf{X}_{it}, M_{it})}{\partial M_{it}}\frac{\exp\{\omega_{it}\}\times\theta}{\widetilde{H}^j(Z_{1,it}, \dots, Z_{S,it})} = P_{M,t}, \tag{4.7}$$

where $\theta \overset{\text{def}}{=} \mathbb{E}[\exp\{\eta_{it}\}|\ \Xi_{it}]$ is a constant. The equation can be transformed to obtain the following nonparametric equation (for the derivation, see the Online Supplementary Material):

$$\ln V_{M,it} = \ln\left[G_M(\mathbf{X}_{it}, M_{it})\theta\right] - \eta_{it}, \tag{4.8}$$

where function $G_M(\mathbf{X}_{it}, M_{it}) \overset{\text{def}}{=} \frac{\partial F(\mathbf{X}_{it}, M_{it})}{\partial M_{it}}\frac{M_{it}}{F(\mathbf{X}_{it}, M_{it})}$ is the elasticity of $F(\cdot)$ with respect to M_{it}, and $V_{M,it} \overset{\text{def}}{=} P_{M,t}M_{it}/\sum_{j=1}^{S} P_{j,t}Y_{j,it}$ is the nominal share of intermediate inputs in total revenue. The share $V_{M,it}$ is observable in the data, and the construction thereof does not require firm-level price data: the information on intermediate expenditures and total revenue would suffice.

Equation (4.8) identifies both the function $G_M(\cdot)\theta$ and the random residual η_{it} on the basis of the following moment conditions for the *ex-post* transitory shock η_{it}:

$$\mathbb{E}[\eta_{it}|\ \mathbf{X}_{it}, M_{it}] = 0. \tag{4.9}$$

Intuitively, (4.8) says that the *unobservable* elasticity of the production function with respect to intermediate inputs can be identified from the *observable* share thereof in total revenue, which

[10] Under the risk neutrality of firms.

must be equal on average (in logs) as implied by the first-order condition. To identify $G_M(\cdot)$ free of constant θ, first note that θ can be identified via $\theta = \mathbb{E}[\exp\{\eta_{it}\}] = \mathbb{E}[\exp\{\ln[G_M(\mathbf{X}_{it}, M_{it})\theta] - \ln V_{M,it}\}]$. Then, the elasticity function is nonparametrically identified as

$$G_M(\mathbf{X}_{it}, M_{it}) = \frac{\exp\{\mathbb{E}[\ln V_{M,it}|\mathbf{X}_{it}, M_{it}]\}}{\mathbb{E}[\exp\{\mathbb{E}[\ln V_{M,it}|\mathbf{X}_{it}, M_{it}] - \ln V_{M,it}\}]}. \tag{4.10}$$

We next identify (up to a constant) the production function $f(\cdot)$ from the already identified function $G_M(\cdot)$ by integrating the ratio of function $G_M(\cdot)$ and M_{it} over M_{it}. Specifically, note that $\frac{G_M(\mathbf{X}_{it}, M_{it})}{M_{it}} = \frac{\partial \ln F(\mathbf{X}_{it}, M_{it})}{\partial M_{it}} \equiv \frac{\partial f(\mathbf{X}_{it}, M_{it})}{\partial M_{it}}$ which, if integrated over M_{it}, yields

$$\int_{\min(M)}^{M_{it}} \frac{G_M(\mathbf{X}_{it}, m)}{m} \mathrm{d}m = f(\mathbf{X}_{it}, M_{it}) + C(\mathbf{X}_{it}), \tag{4.11}$$

where $C(\cdot)$ is the conditional constant of integration. Since the integral can be uniquely obtained via numerical integration using the already identified $G_M(\cdot)$ from (4.8), to identify the production function $f(\cdot)$ from (4.11), we only need to identify $C(\cdot)$. To this end, subtracting (4.11) from (4.3), it is easy to show that the unobservable firm productivity ω_{it} can be expressed as $\omega_{it} = y^*_{it} - \widetilde{h}(\mathbf{Z}_{it}) + C(\mathbf{X}_{it})$ with its first component $y^*_{it} \stackrel{\text{def}}{=} y_{1,it} - \int_{\min(M)}^{M_{it}} \frac{G_M(\mathbf{X}_{it}, m)}{m} \mathrm{d}m - \eta_{it}$ having already been fully identified and thus observable.

Substituting this expression for ω_{it} in the Markov productivity evolution process $\omega_{it} = \psi(\omega_{it-1}) + \zeta_{it}$ in (3.4), we obtain

$$y^*_{it} - \widetilde{h}(\mathbf{Z}_{it}) + C(\mathbf{X}_{it}) = \psi\left[y^*_{it-1} - \widetilde{h}(\mathbf{Z}_{it-1}) + C(\mathbf{X}_{it-1})\right] + \zeta_{it}. \tag{4.12}$$

Under the structural assumptions, all covariates in (4.12) are weakly exogenous because the productivity innovation ζ_{it} is realized at time t after the firm had already predetermined the output ratios (at time $t-1$) and optimized its dynamic inputs (at time $t-b$) for the period t production. That is, $(\mathbf{X}_{it}, \mathbf{Z}_{it}, \mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, y^*_{it-1}) \in \Xi_{it-a}$ for $a > 0$ and all variables can self-instrument. Therefore, the nonparametric equation (4.12) identifies the sum $-\widetilde{h}(\cdot) + C(\cdot)$ (up to a constant) on the basis of the following orthogonality conditions:

$$\mathbb{E}\left[\zeta_{it} \,\middle|\, \mathbf{X}_{it}, \mathbf{Z}_{it}, \mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, y^*_{it-1}\right] = 0. \tag{4.13}$$

Clearly, to identify the output index $\widetilde{h}(\cdot)$ in (4.12), we need to *separably* identify the constant of integration $C(\cdot)$. If successful, that also identifies the production function $f(\cdot)$ in (4.11). The mere fact that $\widetilde{h}(\cdot)$ and $C(\cdot)$ admit non-overlapping arguments alone is insufficient to guarantee their separability because the output ratio decisions and the choice of quasi-fixed inputs are similar in their nature: all are dynamic and exhibit adjustment frictions. In order to distinguish between $\mathbf{Z}_{it}$ and $\mathbf{X}_{it}$, it is imperative that they contain independent variation to rule out the perfect functional dependence between the two. Ackerberg et al. (2015) are the first to raise concerns about this separable unidentifiability problem that many proxy-based estimators suffer from.

In this paper, we use Assumptions 4 and 5(i) about the order of dynamic decisions and the presence of *i.i.d.* adjustment cost shocks to ensure that dynamic inputs $\mathbf{X}_{it} = (K_{it}, N_{it}, L_{it})'$ contain independent variation absent in $\mathbf{Z}_{it}$. Specifically, contrasting

$$Z_{s,it} = Z_{s,it-1} + \mathbb{Z}_{s,t-1}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1}) = Z_{s,it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \omega_{it-1}) \quad \forall s \tag{4.14}$$

with

$$\begin{aligned}
K_{it} &= (1-\delta)K_{it-1} + \mathbb{I}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-1}) = K_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it}, \xi_{it-b}, \omega_{it-1}) = K_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \xi_{it-b}, \omega_{it-1}),\\
N_{it} &= N_{it-1} + \mathbb{N}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-1}) = N_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it}, \xi_{it-b}, \omega_{it-1}) = N_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \xi_{it-b}, \omega_{it-1}),\\
L_{it} &= L_{it-1} + \mathbb{L}_{t-b}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-b}, \xi_{it-b}, \omega_{it-1}) = L_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it}, \xi_{it-b}, \omega_{it-1}) = L_{it}(\mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, \xi_{it-b}, \omega_{it-1}),
\end{aligned}$$

where we have made use of $\mathbf{Z}_{it-b} = \mathbf{Z}_{it}$ in the second equality and substituted (4.14) for $\mathbf{Z}_{it}$ in the last equality, it is evident that all three dynamic inputs contain independent, albeit unobservable, variation that comes from ξ_{it-b} absent in $\{Z_{s,it}\}$ because the former is realized *after* the firm chooses $\mathbf{Z}_{it}$ (recall that $\xi_{it-b} \notin \Xi_{it-1}$). This enables us to separate output index $\widetilde{h}(\cdot)$ from $C(\cdot)$. Thus, with the assumption of a cost shock observable to the firm after it predetermines its relative output specialization, the perfect functional dependence problem is resolved (even if the shock is not directly observed by an econometrician). The output index function $\widetilde{h}(\cdot)$ is separably identified (up to a constant); so is the constant of integration $C(\cdot)$. The autoregressive mean function $\psi(\cdot)$ is identified by "profiling" out the already identified $\widetilde{h}(\cdot)$ and $C(\cdot)$. With $C(\cdot)$ identified, we can identify (up to a constant) function $f(\cdot)$ from (4.11).

Remark 3 Without an explicit conceptual framework describing the firm's decision process about its output mix configuration like the one formalized in Section 3, the proxy-based multi-product model in (4.4) generally *cannot* be identified in the conventional setup, unless some other valid non-input non-output excluded instruments are available. One may think that this paper's framework rendering output ratios self-instrumenting may be forgone in favor of employing lagged outputs/inputs to instrument for outputs on the right-hand side of the equation. However, under the standard assumptions made in the literature, it is easy to show that no lag of output/input is a valid instrument since they all fail to provide any additional *relevant* exogenous variation for the instrumented outputs after conditioning the model on the self-instrumenting variables $(\mathbf{X}_{it}, \mathbf{X}_{it-1}, \mathbf{Z}_{it-1}, M_{it-1})$ already appearing in (4.4). Similar to Gandhi et al.'s (2020) arguments in the context of freely varying inputs, the additional lags would be irrelevant for the output ratios thereby invalidating them as instruments. We tackle this under-identification problem by assuming product-specificity of quasi-fixed inputs and costly cross-product reallocation thereof which renders output ratios dynamic and predetermined and, therefore, self-instrumentable. However, even this is insufficient for identification, necessitating an additional structural assumption about differential information sets available to the firm when making dynamic decisions, which we formalize in terms of the firm's knowledge of factor adjustment cost shocks. Without this additional structure, the multi-product model would be subject to the Ackerberg et al. (2015) critique.

Remark 4 In contrast to available alternatives, our model facilitates a meaningful hurdle-free measurement of the firm's returns to scale. Specifically, the identified multi-product production function in (4.3) has outputs enter the right-hand side of the equation in the *ratio* form ensuring that the estimated input elasticities of the production function measure the effect of input expansion not only on the left-hand-side output but *all outputs proportionately.* The latter is in line with the conventional multi-output definition of returns to scale dated as far back as Panzar & Willig (1977) whereby a proportional change in all inputs is compared with the resultant proportional change in all outputs. More formally, $\text{RTS}_{it} \overset{\text{def}}{=} \frac{\partial \ln \kappa}{\partial \ln \lambda}\big|_{\kappa=\lambda=1}$ for $H(\kappa \mathbf{Y}_{it}) = F(\lambda \mathbf{X}_{it}, \lambda M_{it}) \exp\{\omega_{it}+\eta_{it}\}$, where $\mathbf{Y}_{it}$ is the vector of all outputs (see Färe & Primont, 1995). Making use of the linear homogeneity of the output index function, i.e., $H(\kappa \mathbf{Y}_{it}) = \kappa H(\mathbf{Y}_{it})$, it is easy to show that the RTS measure for the multi-product production function is well-defined in this paper's framework and given by the familiar sum of input elasticities: $\text{RTS}_{it} = \sum_{a\in\{k,n,l,m\}} \frac{\partial f(K_{it},N_{it},L_{it},M_{it})}{\partial a_{it}}$. In particular, from the up-to-a-constant identification of $f(\cdot)$ from (4.11), all gradients of $f(\cdot)$ are identified too, thus identifying the RTS measure. As desired, it quantifies the proportional change in outputs relative to that in inputs.

4.3 Nonparametric Estimation

To operationalize our identification strategy, we estimate the unknown production function $f(\cdot)$, the output index function $\widetilde{h}(\cdot)$ and the (unobservable) firm-level persistent productivity ω_{it} via a two-stage procedure. The estimation is nonparametric. All unknown functions in (4.8) and (4.12) are globally approximated using cubic B-spline sieves, which are a popular type of linear (in parameters) sieves and known for their superior finite-sample performance over other linear sieves, including widely used polynomial series. Essentially, B-splines provide means for a *smooth* piecewise polynomial approximation to the unknown function, with the number and size of subintervals determined via data-driven cross-validation.

Under mild regularity conditions, our nonparametric two-step sieve M-estimator is consistent and asymptotically normal. Furthermore, given the just-identification of our model and so long as we use linear sieves (such as B-splines), one can make use of the numerical equivalence between the consistent estimator of the asymptotic variance of *nonparametric* sieve "parameters" and a consistent estimator of the asymptotic variance derived for these parameter estimators as if the estimated model were actually of a *parametric* form. Thus, in practice, one can use the variance formula for a parametric two-step estimator to consistently estimate the variance of our nonparametric sieve estimator. This streamlines asymptotic inference. Alternatively, one can use bootstrap methods.

We provide a detailed discussion of the estimation, cross-validation and inference procedures in the Online Supplementary Material.

5 Data

The annual farm-level panel data come from the Survey of Account Statistics for Agriculture and Forestry administered by the Norwegian Institute of Bioeconomy Research. The survey includes the information on farm production and economic data collected annually from about 1,000 farms from different regions, farm size classes and types of farms. Participants are selected randomly, and participation in the survey is voluntary.[11] The survey covers a representative sample of agricultural holdings where a substantial share of the family's income is derived from the holding. There is no limit on the number of years a farm may be included in the survey. Approximately 10% of the farms surveyed are replaced every year.

Norway is a marginal growing area for many important crops and, due to the climate, the grain yields per hectare are significantly lower than those in most other countries in Europe. In many regions of Norway, growing fodder crops—mainly grass—is more or less the only alternative. Sales of milk and meat are the two largest sources of income in Norwegian agriculture, amounting to about two thirds of the gross output in 2019 (Knudsen, 2020). To showcase our methodology, we therefore restrict our sample to the farms with combined dairy and beef production. That is, we consider multi-product farms producing two outputs: milk and beef, i.e., $S = 2$. Norwegian combined dairy and beef herds are small in a European context, with the average herd size of 27.8 cows in 2018 (Knudsen, 2020), and thus, the farms that we study are modest in size.

We limit our empirical analysis to the post-2001 period and exclude survey data from the prior years during which dairy farming (one of the two products in our application) in Norway was heavily regulated by the government using a quota-based scheme (e.g., see Jervell & Borgen, 2000). Since then, the Norwegian dairy quota system has been reformed and effectively dismantled by allowing a barrier-free quota trading among individual farms. Quotas are no longer binding in practice, and they essentially remain such in name only (see Vik et al., 2019). Having said that, the continued presence of the milk quota trading system can be a source of adjustment costs for the profit-maximizing farms which, however, can be seamlessly folded into our conceptual framework via the time-varying adjustment costs $\mathrm{C}_t^{\mathcal{Z}}(\cdot)$ associated with the dynamic choice of the farm's relative specialization in milk production using the beef-to-milk ratio.

Given the need for lagged variables in our estimation, we only keep farms for which at least two years of data are available. We also remove the observations that lie in the top and the bottom 1 percentiles of the joint empirical distribution of output and input variables in order to minimize distortionary effects of extreme values on nonparametric estimation. Our operational sample includes 726 farms with a total of 5,037 observations over the 2002–2018 period.

[11] The voluntary participation in the survey may, in theory, give rise to the self-selection issues. Analyzing the extent of their effects on the empirical results as well as attempting to control for them is, however, beyond the scope of our paper. Furthermore, firm sample selection has also become commonly accepted as a *practical* non-issue in the production function estimation literature so long as the data are kept unbalanced. For instance, Levinsohn & Petrin (2003, p.324) write: "The original work by Olley and Pakes devoted significant effort to highlighting the importance of not using an artificially balanced sample (and the selection issues that arise with the balanced sample). They also show once they move to the unbalanced panel, their selection correction does not change their results."

The two output quantities are as follows: Y_1 – milk measured in liters, and Y_2 – beef measured in kilograms. We also make use of the information on total revenue R as an output aggregator commonly used in the single-output production function specifications. The four input quantities are N – land measured in hectares, K – capital that includes farm machinery, buildings and the value of livestock and is measured using the value of fixed assets, L – own and hired farm labor, measured in hours, M – materials that include cost of concentrates, preservatives, medicines, fuel, purchased feed and other costs associated with animal husbandry. Both K and M are the nominal values, deflated to the real 2018 Norwegian Kroner (NOK) using the consumer price index and the price index of other variable costs, respectively.

Table 1 reports summary statistics for our data. Of particular interest are the share of milk sales in total revenue (S_1) and the beef-to-milk output quantity ratio ($Z_2 = Y_2/Y_1$). Their respective empirical distributions are also plotted in Figure 1. As expected, farms in Norway predominantly specialize in the production of milk. Although, with the average milk revenue share of 75% and substantial heterogeneity in the product mix across farms around that average (SD = 12.5%), the data are consistent with beef production playing a non-negligible role. Combined dairy-and-beef farmers have flexibility with the extent of their beef production. A cow has roughly one calf a year and, of course, about half of them are heifers and the other half are bulls. A farmer can choose to keep the bull calf, sell it to other farmers or send it for slaughter. A farmer can also buy additional bull calves in the market. In that sense, not only does the relative specialization between beef and dairy farming vary from farmer to farmer (and year to year) but it also is not a static decision but a dynamic choice subject to natural adjustment frictions.

Overall, Norwegian farming fits our conceptual model well, including the assumption about dynamic adjustments in beef-to-milk specialization, whereby beef is a byproduct of dairy production, as well as the assumption of competitive markets given a high degree of homogeneity of milk and beef products and a small average farm size. The assumption of perfectly competitive markets is very common in productivity analyses of dairy farming. For instance, perfect competition in the output and/or input markets in this context has also been assumed by Tsionas et al. (2015), Latruffe et al. (2017) and Frick & Sauer (2018), to name a few recent examples. More broadly, there is an ongoing debate in the literature about the market structure in agriculture, particularly in the wake of dramatic transformations of the sector in the last decades in both the U.S. and Europe (Sexton, 2013; Sheldon, 2017). The emerging conclusion is that agricultural markets are likely more competitive than their structure might suggest. Furthermore, concerns about rising concentration are usually expressed in the context of food processing and retailing. Given the modest size of Norwegian dairy *producers* (see Table 1) and naturally high product homogeneity, assuming that they exert no significant market power in the product markets and the market for intermediates (e.g., medicines, fuel and feed) is a reasonable first-order approximation.[12]

[12]Recall that the extent to which we need output prices to be exogenous to the firms is essentially limited to the derivation of a static first-order condition with respect to the freely varying intermediate input M_{it}.

6 Results

This section reports nonparametric estimates of the multi-product production technology of dairy-and-beef farms in Norway obtained via our proposed two-stage methodology using cubic B-splines.

In addition to our multi-product model (3.1), we also estimate a more popular single-output specification where the two outputs are *a priori* aggregated into a single total output measure. As is oftentimes done in the literature, here we use the deflated total revenue as an aggregate measure of the farm's total output. To ensure maximal comparability, this single-output model is also estimated nonparametrically in two stages using cubic B-splines. The corresponding estimation procedure closely follows Gandhi et al. (2020) and, essentially, is identical to that of our model except for $\widetilde{h}(\cdot)$ being suppressed in all equations. By estimating this single-output total-revenue model, we seek to empirically assess the sensitivity of the production function estimates to the treatment of multiple outputs in production models.

Both models are cross-validated to select the optimal number of knots used to generate B-spline basis functions.[13] Since our estimation strategy requires the use of lags, the reported results from both models are for the 2003–2018 period.

6.1 Inputs Elasticities

Table 2 summarizes nonparametric estimates of the input elasticities of the production function $\frac{\partial f(K_{it},N_{it},L_{it},M_{it})}{\partial a_{it}}$ $\forall\ a_{it} \in \{k_{it}, n_{it}, l_{it}, m_{it}\}$ from the two models, computed using numerical derivatives of the estimated function $\widehat{f}(\mathbf{X}_{it}, M_{it})$.[14] The empirical results suggest that the total-revenue model produces systematically smaller elasticity estimates with respect to all dynamic inputs: labor, capital and land.[15] This is vividly illustrated in Figure 2 which reports box-plots of observation-specific elasticities over all farm-years based on the two models.

The distributions of input elasticities from the total-revenue model are all shifted downwards compared to those of the estimates from the multi-product model. But the relative differences across the models are the largest in the instance of land input. On average, land elasticity is 0.19 vs. 0.13, with the total-revenue model estimating a smaller contribution of land by roughly a third.

Perhaps, an examination of scale elasticity provides a more intuitive way to compare the empirical implications of the two approaches to modeling the farm production in Norway. Following Remark 3, the returns to scale (or scale elasticity) are computed as the sum of estimated input elasticities. The magnitude less than/equal to/greater than 1 corresponds to decreasing/constant/increasing returns to scale. In the case of our multi-product model, this scale

[13]The data consistently favor the choice of 1 and 2 (equiquantile interior) knots in the first and second stages, respectively. While at first these might appear to imply simple approximations, in practice the implemented sieve approximations are highly flexible and involve hundreds of parameters due to the use of tensor products, as explained in detail in the Online Supplementary Material.

[14]Recall that the sieve approximations are implemented in logs, so the inputs inside the log production function $f(\cdot)$ are logged too. The elasticities are just the first-order gradients of the estimated $\widehat{f}(\cdot)$ with respect to its arguments.

[15]Note that the material elasticity estimates are identical across the two models because they both share the same first-stage "material share equation" (4.8).

elasticity measures the proportional change in *all* outputs relative to that in inputs. When outputs are aggregated using total revenues, the interpretation of equiproportional expansion in all farm outputs however cannot be guaranteed.

The RTS estimates from the two models are summarized in Table 2 as well as in Figure 3 that presents "violin" box-plots of their empirical distributions. Expectedly, the total-revenue model produces notably smaller point estimates of scale elasticity. With the two-sided 95% confidence interval for mean scale elasticity estimated at (0.946, 1.012), our multi-product model, on average, provides statistical evidence in support of *constant* returns to scale in Norwegian farming. In contrast, the total-revenue model suggests *decreasing* returns to scale: the average RTS estimate is 0.857, with the corresponding one-sided 95% upper confidence bound of $0.883 < 1$.

The documented differences in the production function estimates across the two models are material because they lead to the contrasting conclusions about scale efficiency of farming in Norway. While the average RTS estimates from our multi-product model are largely consistent with the expectation that the sector be exhibiting unitary scale elasticity in the long run (and thus, scale efficiency), the evidence from a single-product model generally suggests scale *in*efficiency. That is, using total revenues as an aggregate output measure, one may conclude that the average multi-product farm in Norway may be too large and that there is the potential for cost saving via scaling down the production. The latter is especially puzzling given the small average size of farms in our data.

6.2 Cross-Output Elasticity

One of the advantages of our multi-output model is its ability to identify the production relationship between individual outputs along the farm's production possibilities frontier, which the total-revenue model is unable to deliver given its *a priori* assumption of perfect output substitutability effectively embedded in a linearly additive revenue-based output aggregator. In what follows, we discuss the estimates of this cross-output relationship from our multi-product model.

We first examine the estimates of the cross-output elasticity $\frac{\partial y_{1,it}}{\partial y_{2,it}}$. With the output index function $H(\cdot)$ representing the output frontier and with the farm producing only two outputs, this elasticity can be intuitively interpreted as measuring the *technological* trade-off between the maximal production of two outputs (milk and beef) given the farm's inputs and productivity at the time of production. It quantifies the ease with which, conditional on available inputs and productivity, production can be reconfigured, at the margin, towards either one of the two outputs. Via implicit differentiation of the normalized multi-product production function in (4.3), we can evaluate the cross-output elasticity as $\frac{\partial y_{1,it}}{\partial y_{2,it}} = \frac{\partial h(Z_{2,it})}{\partial z_{2,it}}\big[\frac{\partial h(Z_{2,it})}{\partial z_{2,it}} - 1\big]^{-1}$, where elasticity $\frac{\partial h(Z_{2,it})}{\partial z_{2,it}}$ is obtained using numerical derivatives of the estimated function $\widehat{h}(Z_{2,it})$.

Table 2 reports a summary of the cross-output elasticity estimates. The empirical estimates are negative thereby corroborating the trade-off between the milk and beef production at the statistically significant mean value of –0.294, suggesting that, on average, a 0.3% expansion in milk production can be attained by means of reducing beef products by 1%, *ceteris paribus.* Put differently, the average "shadow price" of a percent expansion in the milk output is a 3.4% (= 1/0.294) reduction in beef production.

The cross-output elasticity is negative and less than one in absolute value for *all* observations, as can be seen from Figure 4 that plots a kernel density of these estimates. The finding of an "inelastic" trade-off between milk and beef is expected given that, for most farms in our sample, milk is the main source of revenue and the beef production to some extent is a "byproduct" of dairy farming, although the exact beef-to-milk specialization is at the farmer's discretion. Furthermore, as can also be observed from Figure 5, the more specialized in milk a farm is, the higher the shadow price of an additional expansion of milk production is in terms of the required reduction in beef products.[16] This is expected since, typically, a highly specialized dairy farmer will not normally have much space available for extra bulls in the cowshed.

When interpreting the cross-output elasticities, it is imperative to remember that, unlike in a simple static setup where the firm chooses an output mix such that the relative market output price equals the corresponding technological "price" on a boundary of the producible output set, the $\frac{\partial y_{1,it}}{\partial y_{2,it}}$ estimates reflect the multitude of factors influencing farms' production decisions. This happens because, owing to product specificity of inputs and a costly cross-product reallocation thereof, the farm's output mix is chosen in advance via the multi-period dynamic optimization. Hence, the *realized* cross-output elasticity will likely differ from the contemporaneous output price ratio because the former depends on the firm's expectations of output prices as well as on (potentially firm-specific) frictions associated with adjustments in the product mix and dynamic inputs.

Each cross-output elasticity estimate discussed above corresponds to a single realized point on the frontier of a producible output set for each farm-year which impedes a clean *ceteris paribus* analysis of the milk-vs-beef technological relationship in Norwegian farming. To circumvent this problem, we construct the "counterfactual" production possibilities frontier $\{(Y_1^*, Y_2^*) : \widehat{H}(Y_1^*, Y_2^*) = H_0\}$ for a representative median farm such that $H_0 = \text{Med}[\exp\{\widehat{f}(\mathbf{X}_{it}, M_{it}) + \widehat{\omega}_{it} + \widehat{\eta_{it}}\}]$, where the output index $\widehat{H}(\cdot)$ is constructed using $\widehat{H}(Y_1^*, Y_2^*) = \exp\{\ln Y_1^* + \widehat{h}(Z_2^*)\}$ with $Z_2^* = Y_2^*/Y_1^*$. Figure 6 presents the graph of this output frontier (solid line) constructed around the median values of output quantities observed in the data.

As expected, the estimated frontier of the output set is downward-sloping and concave to the origin. To accentuate its concavity, we also plot the *linear* production possibilities frontier (long-dashed line) implied by the revenue-based aggregation of the two outputs. We construct the latter using average output prices and, to ensure maximal comparability, the same H_0 value from our multi-product model.[17] Evidently, by explicitly estimating the cross-output technological relationship as opposed to *a priori* assuming perfect substitutability between the two, we are able to detect

[16] Recall that $\left|\frac{\partial y_{1,it}}{\partial y_{2,it}}\right|$ of lower magnitude implies a higher "shadow price" of an additional 1% expansion in Y_1 as measured in percents of the technologically necessitated reduction in Y_2.

[17] The latter means that the relative position of the two frontiers is artificial and, hence, their slopes on the graph are not comparable for any fixed output quantities.

convexity of the farm's output set which is indicative of technological complementarities in the production of milk and beef. Such complementarities are expected given that both products involve raising cattle and because the beef production, to an extent, is a byproduct of milk production (although the degree of specialization in beef products remains a farmer's choice). The revenue-based aggregation however unrealistically assumes such technological complementarities away in favor of perfect substitutability between cow milk and beef.

On a final note, we also consider the estimates of farm productivity. Given that the multi- and single-product models rely on inherently different aggregations of outputs [$H(\cdot)$ vs. $R(\cdot)$] and hence produce different estimates of the production function, the estimates of ω_{it} from these two models are not comparable. We therefore focus on the farm productivity estimates from our multi-product model only. Since it is not the primary focus of the paper, discussion of these productivity results is relegated to the Online Supplementary Material.

7 Conclusion

This paper contributes to the literature by providing a new methodology for proxy variable identification of multi-product production technologies which not only does not require the access to information beyond that available in most datasets but also successfully identifies the cross-output technological relationship within the firm by virtue of not *a priori* aggregating outputs. We consider the estimation of (single-equation) multi-product production functions where outputs form an unspecified aggregator index which represents a set of maximal quantities of all outputs producible from inputs given the firm's productivity.

We show that, in the case of multi-product firms, the conventional assumptions alone are not sufficient to identify the production function. To identify the multi-product model, we lay down a set of additional structural assumptions about the firm's technology, economic environment and its decision-making process. Building on this framework, we tackle the endogenous static inputs in the presence of latent firm productivity by building upon Gandhi et al. (2020) whose system approach we extend to the case of multiple outputs.

There are several promising extensions of our model worth exploring in future research. For instance, one of the key components of our identification strategy comes from the information contained in the firm's static optimality condition, which we derive under the assumption that the markets are perfectly competitive. Because this market structure assumption is not suitable for all empirical applications, one practically important extension of our model would be focused on allowing for monopolistic power by firms. Relying on the returns-to-scale restrictions along the lines of Flynn et al. (2019) might be a promising way of achieving identification of the production technology in this case. Second, in our conceptual model, we implicitly assume that all firms are multi-product (i.e., there are no fully specialized single-product firms) and that each of them produces all outputs. However, researchers may also be interested in analyzing industries in which firms produce different product portfolios or for which the datasets are a mix of single- and multi-output firms. A richer framework suitable for that kind of analysis will include an explicit formulation of the firm's endogenous choice of the scope of its products and not just the choice of the relative composition of outputs within a fixed product mix as we currently do in this paper. We leave these extensions for future research.

References

Ackerberg, D. A., Benkard, C. L., Berry, S., & Pakes, A. (2007). Econometric tools for analyzing market outcomes. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of Econometrics*, volume 6A. North Holland.

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, *83*, 2411—2451.

Amsler, C., Prokhorov, A., & Schmidt, P. (2016). Endogeneity in stochastic frontier models. *Journal of Econometrics*, *190*, 280–288.

Amsler, C., Prokhorov, A., & Schmidt, P. (2017). Endogenous environmental variables in stochastic frontier models. *Journal of Econometrics*, *199*, 131–140.

Atkinson, S. E., Cornwell, C., & Honerkamp, O. (2003). Measuring and decomposing productivity change: Stochastic distance function estimation versus data envelopment analysis. *Journal of Business and Economics Statistics*, *21*(2), 284–294.

Atkinson, S. E. & Primont, D. (2002). Stochastic estimation of firm technology, inefficiency, and productivity growth using shadow cost and distance functions. *Journal of Econometrics*, *108*, 203–225.

Atkinson, S. E. & Tsionas, E. G. (2016). Directional distance functions: Optimal endogenous instruments. *Journal of Econometrics*, *190*(2), 301–314.

Bernard, A. B., Redding, S. J., & Schott, P. K. (2010). Multi-product firms and product switching. *American Economic Review*, *100*, 70–97.

Caballero, R. J., Engel, E. M. R. A., & Haltiwanger, J. (1997). Aggregate employment dynamics: Building from microeconomic evidence. *American Economic Review*, *87*, 115–137.

Charnes, A., Cooper, W. W., Lewin, A. Y., & Seiford, L. M. (2013). *Data Envelopment Analysis: Theory, Methodology, and Applications.* Springer Science + Business Media.

Cooper, R. & Willis, J. (2003). The cost of labor adjustment: Inference from the gap. NBER Working Paper 10006.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica*, *79*, 1407–1451.

De Loecker, J., Eeckhout, J., & Unger, G. (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics*, *135*, 561–644.

De Loecker, J., Goldberg, P. K., Khandelwal, A. K., & Pavcnik, N. (2016). Prices, markups, and trade reform. *Econometrica*, *84*, 445–510.

De Loecker, J. & Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, *102*, 2437–3471.

Dhyne, E., Petrin, A., Smeets, V., & Warzynski, F. (2016). Multi-product firms, import competition, and the evolution of firm-product technical efficiencies. Working Paper, Aarhus University.

Färe, R. & Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications.* Boston, MA: Kluwer Academic Publishers.

Fernández, C., Koop, G., & Steel, M. F. J. (2000). A Bayesian analysis of multiple-output production frontiers. *Journal of Econometrics*, *98*, 47–79.

Fernández, C., Koop, G., & Steel, M. F. J. (2002). Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture. *Journal of the American Statistical Association*, *97*(458), 432–442.

Fernández, C., Koop, G., & Steel, M. F. J. (2005). Alternative efficiency measures for multiple-output production. *Journal of Econometrics*, *126*, 411–444.

Flynn, Z., Gandhi, A., & Traina, J. (2019). Measuring markups with production data. Working Paper, University of Chicago.

Foreman, L. (2014). Characteristics and production costs of U.S. corn farms, including organic, 2010. USDA ERS Economic Information Bulletin. Number 128.

Frick, F. & Sauer, J. (2018). Deregulation and productivity: Empirical evidence on dairy production. *American Journal of Agricultural Economics*, *100*, 354–378.

Gandhi, A., Navarro, S., & Rivers, D. (2020). On the identification of gross output production functions. *Journal of Political Economy*, *128*, 2973–3016.

Goldberg, P. K., Khandelwal, A. K., Pavcnik, N., & Topalova, P. (2010). Multiproduct firms and product turnover in the developing world: Evidence from India. *Review of Economics and Statistics*, *92*, 1042–1049.

Greene, W. H. (2008). The econometric approach to efficiency analysis. In H. O. Fried, C. A. K. Lovell, & S. S. Schmidt (Eds.), *The Measurement of Productive Efficiency and Productivity Change*. Oxford University Press.

Grieco, P. L. E. & McDevitt, R. C. (2017). Productivity and quality in health care: Evidence from the dialysis industry. *Review of Economic Studies*, *84*, 1071–1105.

Griffiths, W. E. & Hajargasht, G. (2016). Some models of stochastic frontiers with endogeneity. *Journal of Econometrics*, *190*, 341–348.

Griliches, Z. & Mairesse, J. (1998). Production functions: The search for identification. In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium* (pp. 169–203). Cambridge University Press.

Jervell, A. M. & Borgen, S. O. (2000). Distribution of dairy production rights through quotas: The Norwegian case. In H. K. Schwarzweller & A. P. Davidson (Eds.), *Dairy Industry Restructuring*. New York: Elsevier Science Inc.

Knudsen, H. (2020). Norwegian agriculture: Status and trends 2019. *NIBIO POP*, *6*.

Kumbhakar, S. & Lovell, C. A. K. (2000). *Stochastic Frontier Analysis*. New York, NY: Cambridge University Press.

Kumbhakar, S. C. (2013). Specification and estimation of multiple output technologies: A primal approach. *European Journal of Operational Research*, *231*, 465–473.

Kumbhakar, S. C. & Tsionas, E. G. (2016). The good, the bad and the technology: Endogeneity in environmental production models. *Journal of Econometrics*, *190*, 315–327.

Kuosmanen, T., Johnson, A., & Saastamoinen, A. (2015). Stochastic nonparametric approach to efficiency analysis: A unified framework. In J. Zhu (Ed.), *Data Envelopment Analysis*. Springer Science + Business Media.

Kutlu (2010). Battese-Coelli estimator with endogenous regressors. *Economics Letters*, *109*, 79–81.

Latruffe, L., Bravo-Ureta, B. E., Carpentier, A., Desjeux, Y., & Moreira, V. (2017). Subsidies and technical efficiency in agriculture: Evidence from European dairy farms. *American Journal of Agricultural Economics*, *99*, 783–799.

Levinsohn, J. & Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, *70*, 317–341.

Maasoumi, E. & Nickelsburg, G. (1988). Multivariate measures of well-being and an analysis of inequality in the Michigan data. *Journal of Business and Economic Statistics*, *6*, 327–334.

Maasoumi, E. & Racine, J. S. (2016). A solution to aggregation and an application to multidimensional 'well-being' frontiers. *Journal of Econometrics*, *191*, 347–383.

Malikov, E., Zhao, S., & Kumbhakar, S. C. (2020). Estimation of firm-level productivity in the presence of exports: Evidence from China's manufacturing. *Journal of Applied Econometrics*, *35*(4), 457–480.

Marschak, J. & Andrews, W. H. (1944). Random simultaneous equations and the theory of production. *Econometrica*, *12*, 143–205.

Mensah, A. & Brümmer, B. (2016). A multi-output production efficiency analysis of commercial banana farms in the Volta region of Ghana: A stochastic distance function approach. *African Journal of Agricultural and Resource Economics*, *11*, 305–316.

Newey, W. K. & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, *71*(5), 1565–1578.

O'Donnell, C. J. (2014). Econometric estimation of distance functions and associated measures of productivity and efficiency change. *Journal of Productivity Analysis*, *41*, 187–200.

O'Donnell, C. J. & Nguyen, K. (2013). An econometric approach to estimating support prices and measures of productivity change in public hospitals. *Journal of Productivity Analysis*, *40*, 323–335.

Olley, G. S. & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, *64*, 1263–1297.

Panzar, J. C. & Willig, R. D. (1977). Economies of scale in multi-output production. *Quarterly Journal of Economics*, *91*, 481–493.

Ray, S. (2004). *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*. Cambridge University Press.

Sexton, R. J. (2013). Market power, misconceptions, and modern agricultural markets. *American Journal of Agricultural Economics*, *95*, 209–219.

Sheldon, I. M. (2017). The competitiveness of agricultural product and input markets: A review and synthesis of recent research. *Journal of Agricultural and Applied Economics*, *49*, 1–44.

Tran, K. C. & Tsionas, E. G. (2013). GMM estimation of stochastic frontier model with endogenous regressors. *Economics Letters*, *2013*, 233–236.

Tran, K. C. & Tsionas, E. G. (2015). Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters*, *2013*, 233–236.

Tsionas, E. G., Kumbhakar, S. C., & Malikov, E. (2015). Estimation of input distance functions: A system approach. *American Journal of Agricultural Economics*, *97*, 1478–1493.

Valmari, N. (2016). Estimating production functions of multiproduct firms. Working Paper, Research Institute of the Finnish Economy.

Vik, J., Stæte, E. P., Hansen, B. G., & Nærland, T. (2019). The political robot – The structural consequences of automated milking systems (AMS) in Norway. *NJAS – Wageningen Journal of Life Sciences*, *90–91*. Article 100305.

Table 1. Data Summary Statistics

Variable	Description and Units	Mean	SD	2.5%	Median	97.5%
Y_1	Milk, *liters*	162,917.9	87,243.7	55,816.6	137,924.0	395,184.9
Y_2	Beef, *kilograms*	5,342.4	4,259.9	935.8	4098.0	17,114.5
Z_2	Beef-to-Milk Output Ratio, *kg/l*	0.033	0.020	0.009	0.03	0.083
R	Total Revenue, *2018 NOK*	1,396,520	749,004	487,975	1,192,358	3,342,925
S_1	Milk Revenue Share, *proportion*	0.750	0.125	0.464	0.765	0.929
V_M	Material Revenue Share, *proportion*	0.437	0.086	0.295	0.429	0.632
K	Capital, *2018 NOK*	3,088,791	2,487,961	649,106	2,183,474	10,191,405
N	Land, *hectares (ha)*	35.1	17.9	13.1	31.0	81.2
L	Labor, *hours (hr)*	3,595.5	912.3	2,150	3,472	5,770
M	Intermediate Inputs, *2018 NOK*	625,742	325,080	225,969	550,597	1,449,118

SD is the sample standard deviation. 2.5% and 97.5% are the 2.5th and 97.5th percentiles of the empirical distribution, respectively. The average exchange rate in 2018 was 1 USD for 8.142 NOK.

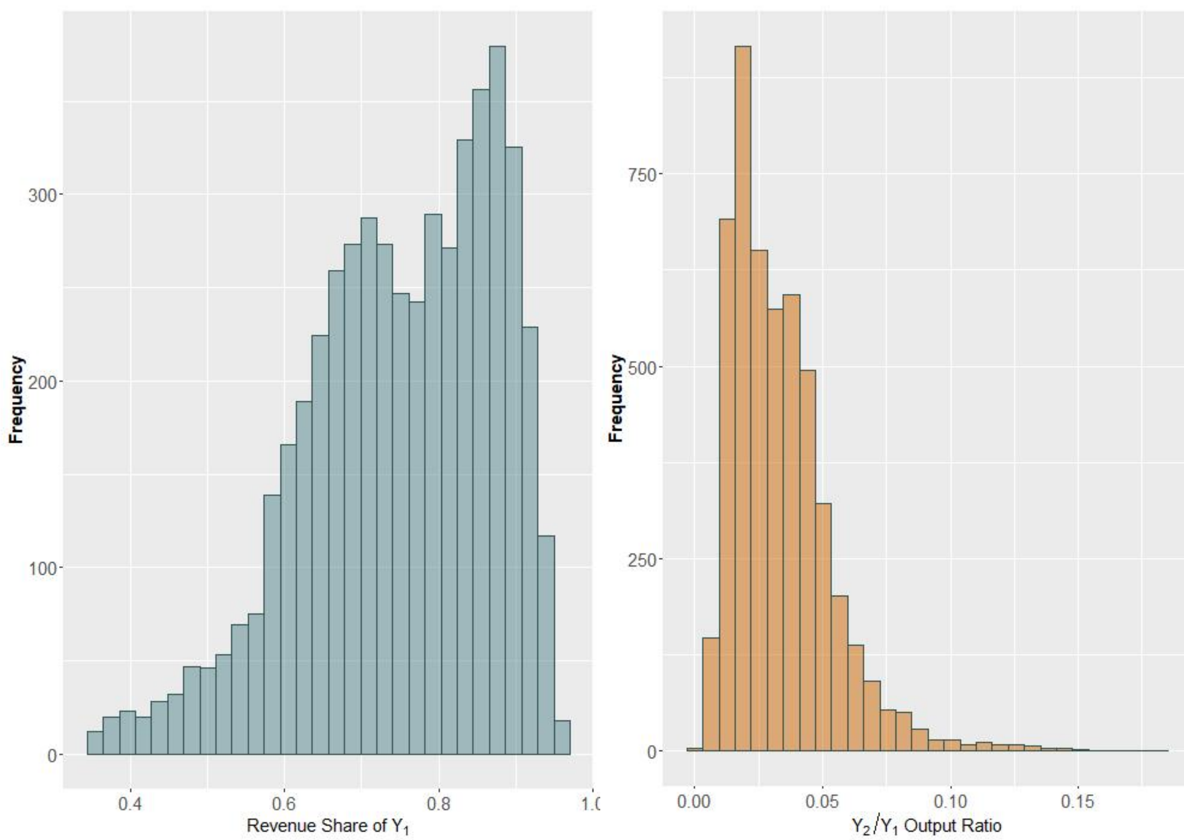

Figure 1. Distribution of the Milk Revenue Share and the Beef-to-Milk Output Quantity Ratio

Table 2. Production Function Estimates

	Multiple Outputs				*Total Revenue*			
	Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
Materials Elasticity	0.438	0.406	0.441	0.453	0.438	0.406	0.441	0.453
	(0.001)				(0.001)			
Labor Elasticity	0.235	0.025	0.246	0.431	0.203	0.024	0.214	0.402
	(0.014)				(0.014)			
Capital Elasticity	0.116	0.021	0.133	0.229	0.083	0.010	0.110	0.187
	(0.005)				(0.005)			
Land Elasticity	0.190	0.012	0.153	0.365	0.133	–0.014	0.109	0.265
	(0.009)				(0.008)			
Scale Elasticity	0.979	0.681	0.959	1.256	0.857	0.628	0.857	1.102
	(0.017)				(0.016)			
Cross-Output Elasticity	–0.294	–0.431	–0.241	–0.128				
	(0.003)							

Material elasticity estimates are expectedly identical across the two models because both share the same first-stage "material share equation." Q1 and Q3 are the first and third quartiles of the empirical distribution, respectively. Bootstrap standard errors in parentheses.

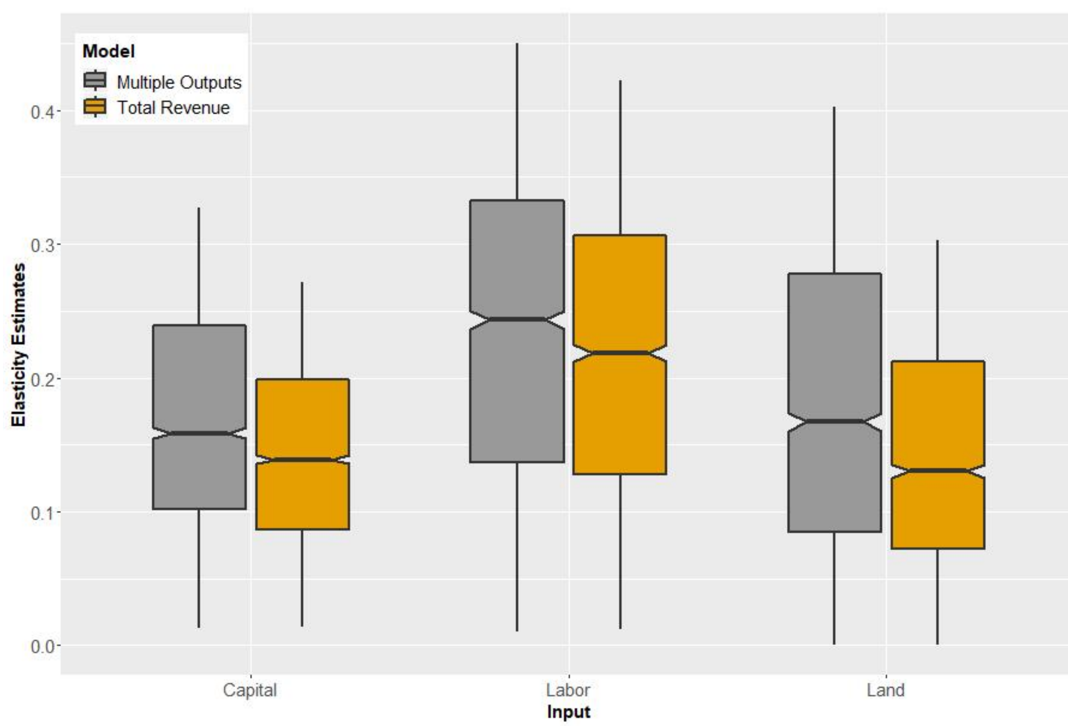

Figure 2. Input Elasticity Estimates

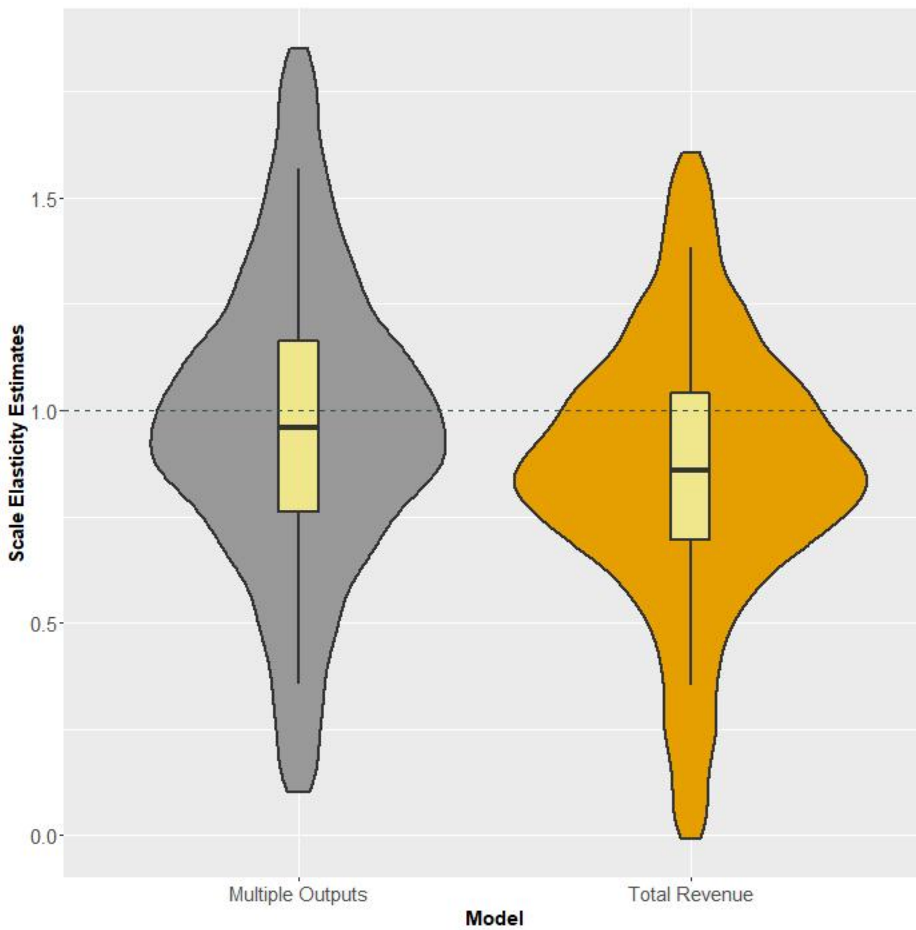

Figure 3. Scale Elasticity Estimates

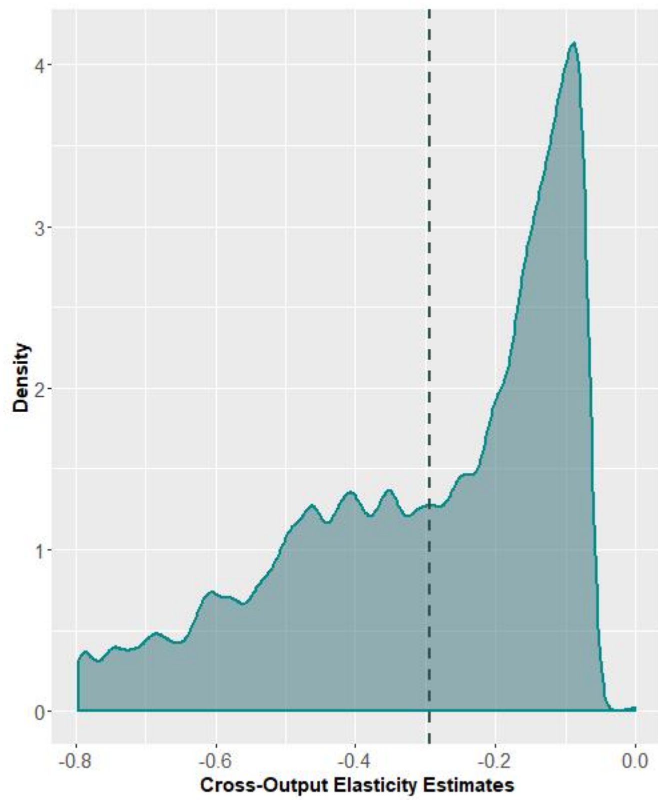

Figure 4. Cross-Output Elasticity Estimates

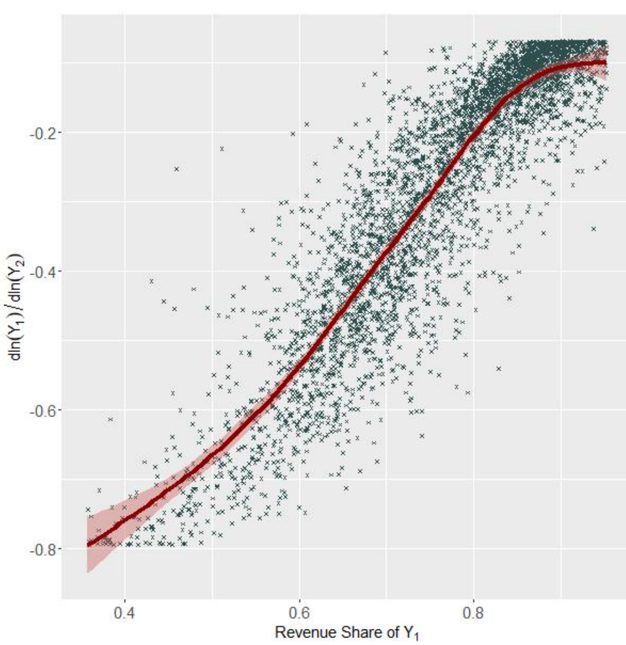

Figure 5. Cross-Output Elasticity Estimates vs. Specialization in Milk Production
[The solid red line is the fitted spline regression.]

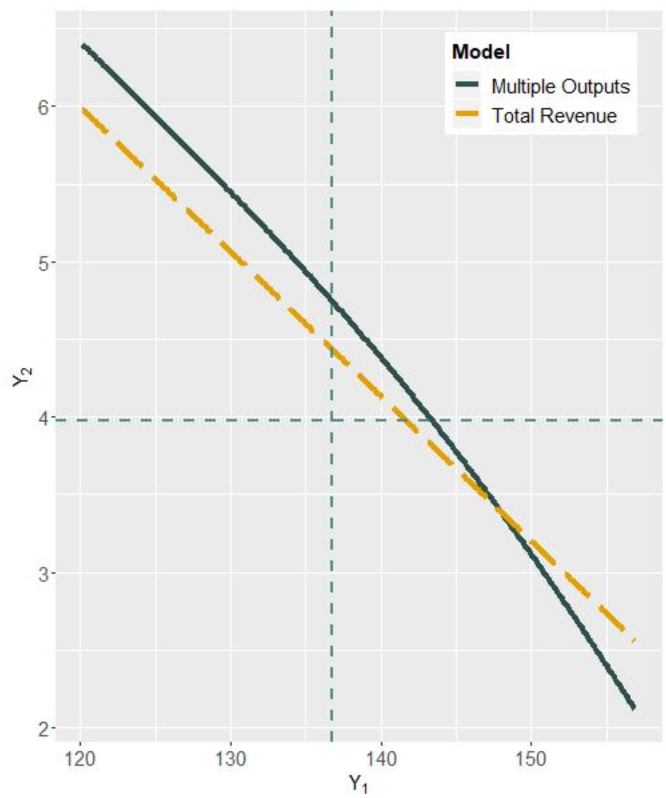

Figure 6. Production Possibilities Frontier at the Median
[Output quantities are in 1,000s. Dashed vertical and horizontal lines correspond to the median output quantities observed in the data.]