Measuring Productivity and Inefficiency Without Quantitative Output Data

Erik Mellander

The Industrial Institute for
Economic and Social Research

is an independent non-profit research institution, founded in 1939 by the Swedish Employers' Confederation and the Federation of Swedish Industries.

Objectives

To carry out research into economic and social conditions of importance for industrial development in Sweden.

Activities

The greater part of the Institute's work is devoted to long-term problems, especially to long-term changes in the structure of the Swedish economy particularly within manufacturing industry.

Board

Address

Industriens Utredningsinstitut
Box 5501, S-114 85 Stockholm, Sweden
Tel. 08-783 80 00

Measuring Productivity and Inefficiency Without Quantitative Output Data

Erik Mellander

Distribution: Almqvist & Wiksell International,
Stockholm, Sweden

IUI Dissertation Series, No. 5
ISSN 1102-6057
ISBN 91-7204-426-8
Printed by AB Grafiska Gruppen
Stockholm 1993

Foreword

The analyses of production, public sector economics and the development of econometric method have a long tradition at the Industrial Institute for Economic and Social Research (IUI). This theoretical study by Erik Mellander combines these areas in an innovative way.

The possibilities of estimating efficiency and productivity are analyzed for situations when reliable measures of production cannot be constructed. This is often the case for public services and also for private production with heterogeneous product qualities. It is demonstrated for a large class of production technologies that data on output are not required to study productivity and efficiency. An econometric model is also developed to implement the results empirically.

This book has been submitted as a Ph.D. thesis at the University of Uppsala. It is the 47th doctoral or licentiate dissertation completed at the Institute since its foundation in 1939. IUI would like to thank Peter Englund and Bertil Holmlund who acted as thesis advisors and Ernst Berndt for his keen interest in, and support of, Mellander's work during his frequent visits to the institute. IUI would also like to acknowledge the importance of two of its former researchers in the early part of this study, namely, Bengt-Christer Ysander, who was Mellander's first thesis advisor and Leif Jansson, who was his collaborator in earlier work on related problems. The generous financial support from the Bank of Sweden Tercentenary Foundation, the Royal Swedish Academy of Sciences and the Sweden-America Foundation is gratefully acknowledged.

Stockholm in March 1993

Gunnar Eliasson

MEASURING PRODUCTIVITY AND INEFFICIENCY WITHOUT QUANTITATIVE OUTPUT DATA

Erik Mellander

Dissertation for the degree of Doctor of Philosophy, Uppsala University, April, 1993.

ABSTRACT

This dissertation investigates the possibilities to estimate the development of productivity and inefficiency in the absence of quantitative measures of the production result.

Chapter I discusses the relevancy of the methods suggested in the thesis and puts them into perspective by relating them to other, more traditional, methods. The basic assumption in the succeeding chapters – that the production technology is homothetic – is considered and the concept of homotheticity is characterized.

In Chapter II a general method is described which allows a production activity to be analyzed by means of input data only. According to duality theory, the input cost shares can be completely specified without any information about output if the technology is homothetic. It is demonstrated that these cost shares can yield information about elasticities of substitution and factor demand and on productivity growth. Effects of returns to scale can not be analyzed, however. Finally, the system of share equations is generalized to allow for technical and allocative inefficiency and it is shown how to compute the effects of these inefficiencies on total cost and input demands.

The purpose of Chapter III is to show that the parameters of the model in the preceding chapter can be identified and, hence, econometrically estimated. It is assumed that the production technology can be characterized by means of a translog cost function and, for simplicity, the attention is confined to the case with (at most) three inputs. Both allocative and technical inefficiencies are modeled parametrically. It is proved that if the disturbances of the cost share equations are joint normally distributed then the information matrix of the Full Information Maximum Likelihood estimator has full rank, which implies that the parameters are (locally) identified.

Chapter IV extends the results to the case when a *value* measure of output is available. This situation is typical of many private service industries. The output market is allowed to be noncompetitive, and the potential markup is assumed to be either known or constant. It is shown that if average cost is strictly increasing in the volume of output then the given data are equivalent to complete information, provided that the markup is known. Accordingly, return to scale can be analyzed and, hence, productivity growth can be completely determined. If the markup is unknown, the results continue to hold conditional on the unknown markup.

Erik Mellander, the Industrial Institute for Economic and Social Research, Box 5501, S – 114 85 Stockholm, Sweden.

Contents

Acknowledgements

The first, and most important, step in the writing of a dissertation is to formulate a suitable research problem. I have my friend, fellow-worker, and thesis advisor, the late Bengt-Christer Ysander to thank for taking care of this part of the work for me.

Bengt-Christer suspected that the low rates of productivity growth reported in studies of public services might be due to the use of inappropriate output measures. The seemingly insuperable difficulties associated with quantifying all the various quality aspects of services made him start thinking about the possibilities to avoid the output measurement problem altogether. This rather bold idea later became the subject of my thesis.

Chapter II is a result of joint work with Bengt-Christer. Thanks are due to the following persons for providing detailed comments on this chapter: Ernst Berndt, Rolf Färe, Lennart Hjalmarsson, Eilev Jansen, Olle Mellander, Joan Muysken, Ishaq Nadiri and Henry Tulkens. The financial support from the Bank of Sweden Tercentenary Foundation is also gratefully acknowledged.

Sune Karlsson did an outstanding job in scrutinizing Chapter III. In addition to pointing out errors and suggesting corrections and improvements he also took time to check my endeavors to solve the problems he had found. Thanks is due also to Harald Lang for a most enlightening discussion.

Chapter IV was much improved by the constructive comments from Fabio Schiantarelli and Thomas Lindh.

The work on this thesis has been carried out at the Industrial Institute for Economic and Social Research (IUI). Over the years, Gunnar Eliasson has been very successful in making the IUI a stimulating research environment, much reflecting his own enthusiasm and open-mindedness. I am also grateful to Gunnar for his interest in and support of my research.

At the IUI, several people have contributed with helpful comments and suggestions. Particular thanks are due to Thomas Andersson, Eugenia Kazamaki Ottersten, Per Lundborg, Sten Nyberg and Pavel Pelikan. I am also grateful to Torsten Dahlquist for helping me to compute numerical examples when I got stuck in my theoretical arguments. Marie-Louise Falk and Maria Hedström kindly and efficiently helped me with a number of practical matters. Furthermore, by handling the printing process, Maria saw to it that my manuscript eventually turned into a thesis.

In his own special way, the late Leif Jansson, formerly at the IUI, greatly contributed to this thesis. I learned a lot from Leif about production theory and econometrics. Joyce Dargay was another important source of inspiration.

Thanks to grants from the Sweden-America Foundation and the Royal Swedish Academy of Sciences I had the opportunity to spend the latter half of 1990 and the beginning of 1991 as Special Student at MIT, Cambridge, Massachusetts. My family and I are sincerely grateful to Ernst Berndt for all his help in making this visit possible and for his generosity to us during our stay. Ernst Berndt also provided me with an invitation to the NBER Summer Institute Productivity Seminars, which gave me several useful ideas. At MIT, I was lucky enough to become a fellow-student of Patrik Edsparr. To see Patrik's extraordinary problem-solving talent at work was truly instructive. Finally, a very special thanks to the Bianconi family, our Cambridge neighbors, for creating such a pleasant atmosphere.

Being a graduate student at the Economics Department in Uppsala has been a very rewarding experience. The seminars and numerous discussions with Gudmundur Gunnarsson, Bengt Hansson and Thomas Lindh and other members of the Growth Study Group have been most beneficial to me. The friendly and skilfull assistance of the administrative staff, represented by Eva Holst, Monica Ekström and Berit Levin, is also gratefully acknowledged.

When Bengt-Christer Ysander's illness forced him to resign, Peter Englund took over the role as my thesis advisor. I certainly could not have wished for a better successor. Peter's comments on my drafts were always constructive and insightful and his encouragement made the work much easier, as did his help on matters not directly related to the thesis. When Peter was on sabbatical leave in the U.S., Bertil Holmlund acted as my advisor, a task which he performed with his usual efficiency and kindness.

Without the encouragement and assistance from my whole family, this thesis may have remained unfinished. To my wife Ann and my daughter Danielle I wish to express my gratitude for all their support, patience and understanding.

Stockholm in March 1993

Erik Mellander

CHAPTER I

Introduction and summary

To try to measure productivity and inefficiency when what is produced (the output) cannot be observed may not seem like a very good idea. It *is* possible, however, and the primary purpose of this thesis is to show how it can be done. While it is purely theoretical, the analysis here is geared directly towards empirical applications.

Even though the subject of the thesis may be interesting in its own right, it is reasonable to ask if it is also relevant. Are there situations when it is necessary or desirable to measure productivity and inefficiency without having quantitative output data? After all, it is hard to conceive of production processes for which no quantitative information can be obtained about the production result. And, if such information were available, why would one ignore it?

In this introduction, the emphasis will be on the relevancy issue. Firstly, it will be shown that there are indeed circumstances requiring the use of the methods suggested in the ensuing chapters. Secondly, it will be argued that in many cases it may be desirable to use them, even when it is not strictly necessary. However, it should be noted that this rides on the assumption that the output volume data are measured with error. While the methodology is completely general, in the sense that it can be applied to any production activity, applications to service production will be stressed because, as a rule, output measurement problems are more severe in service production than in goods production.

First, a few words will be said about the assumptions made in the thesis about the data available for analysis. In this context, the system of national accounts will be adduced as an example of a data source which for some types of production can make it necessary to measure productivity and inefficiency without quantitative output data. Next, the traditional approach to handling output data subject to error will be reviewed. The alternative approach suggested in this thesis will then be outlined in the last two sections.

1. Basic assumptions about data

To focus the analysis on problems associated with output, it is assumed that all relevant information is available about the factors of production (the inputs) and, furthermore, that it is known that this information contains no errors.[1] The available quantitative information about output, on the other hand, is assumed to be defective in some sense and the nature of the defect is assumed to be unknown. In a stylized form, these two assumptions describe the following very common situation: there are strong reasons to suspect that the output measure(s) are inaccurate but information cannot be obtained about the precise character of the errors involved.

The designation defective output data is here taken to include all output data sets except those completely free of error. Thus, on the one extreme, defective output data may be output data subject to an independently and identically distributed random error with zero mean. For simplicity, such errors will be denoted white noise, in accordance with the jargon used in the

[1] Of course, there are problems in the context of input measurement, too. There is, however, a large literature dealing with how these problems can be handled. See Usher (1980) and Hamermesh (1986), for example, for discussions about problems associated with the measurement of capital and labor, respectively.

statistical time series literature. On the other extreme, defective output data will simply be non-existent output data.

It should be noted that the possibility of there being non-quantitative information about output is not excluded here. Specifically, the usefulness of output *value* data is considered in the final chapter of the thesis where it is assumed that, as with input data, the value data are known to contain no errors.

While instances where there is no output information whatsoever are very unlikely, it may be that information exists but cannot be used for the particular purposes of measuring productivity and inefficiency. This point can be illustrated by the treatment of the production of services in the national accounts statistics, especially public services but also several types of private services, e.g. financial services. By convention, the output of the public sector is in the national accounts equalized to the volume of resources used to produce it. This implies *inter alia* that public sector productivity growth is set equal to zero a priori and that inefficiency in the form of excessive factor usage is assumed not to exist.[2] Thus, by construction, these output measures incorporate assumptions about the phenomena to be investigated, i.e. assumptions about productivity and efficiency, and hence obviously cannot be used in analyses of these issues. Concerning output measures for private services which suffer from similar problems, the most well-known examples are given by the banking and insurance industries. For example, in the Swedish National Accounts the (volume of) value added in these sectors is estimated by assuming that average labor productivity increases by a constant yearly rate of 2%. Procedures of this kind are employed in other countries, too.

When quantitative output data are either non-existent or inadequate in

[2] A more extensive discussion of the properties of the public sector output measure used in the national accounts can be found in Chapter II.

the sense just described, then it is necessary to resort to procedures for measuring productivity and efficiency which, like the methods suggested in this thesis, do not require any quantitative output information at all. However, in general some kind of quantitative output information is available in which case it is also possible to use traditional approaches.

2. The traditional approach: output measurement models

The traditional approach to handling the problem of defective output data is to complement the structural model used in the analysis – say a production function or a cost function – by a measurement model that describes the relationship between the available output proxy variable(s) and the unobserved "true" output. Of course, for this to be a meaningful undertaking output data must be neither non-existent nor inadequate in the sense described in the previous section. Accordingly, that is assumed here.

In the simplest possible case the true output is equal to the sum of a single proxy variable and a white noise error. In this case of a purely nonsystematic error there is really no problem and therefore no need to formulate the measurement model explicitly.[3] A benign interpretation of the many studies in which output measurement problems are acknowledged but no attempt is made to take them explicitly into account in the analysis is that, implicitly, they have assumed this particular measurement model. Implicit measurement models are also employed in contexts where output is taken to

---

[3] If the structural model is a production function then it is obvious that the output measurement problem can be neglected; the output measurement error will simply be added to the random error in the structural model. If a cost function is used, then it must be assumed that the output error is not correlated with the error in the cost function.

be multi-dimensional. In that case, one procedure is to aggregate several proxies into an output indicator index. Another technique is to model production as a multiple-output process, using one proxy variable for each output dimension.[4]

There are very few examples of analyses where the measurement model has been formulated explicitly.[5] One of the first to appear was the study by Spady and Friedlaender (1978) of the American trucking industry where the unobserved "true" output was modeled as a function of the traditional output measure – ton-miles – and various other characteristics like average shipment size, average length of haul, insurance coverage, etc. By substituting this function into the structural model (the cost function) Spady and Friedlaender were able to estimate the unobserved output variable and the cost function simultaneously. Comparing the results of this model with those obtained using ton-miles as the output variable they found substantial differences, indicating that the systematic measurement error associated with the ton-miles variable was of considerable practical importance.[6]

---

[4] Examples of studies employing the output indicator index technique are Pauly (1978) and Parsons, Gotlieb and Denny (1992). Analyses where production is modeled as a multiple-output process have become very popular recently; see Grosskopf and Valdmanis (1987), Mester (1987), Eakin and Kniesner (1988), Gyapong and Gyimah-Brempong (1988) and Bernstein (1992).

[5] It may be noticed that the situation is quite different in applied sociology and applied psychology. There, the use of explicit measurement models has a long tradition and software has been developed for estimation of models containing both a structural submodel and a measurement submodel; one example is the LISREL program, a description of which can be found in Jöreskog and Sörbom (1989).

[6] A related approach has recently been used in studies of the production of child care [McKay (1988), Mukerjee, Witte and Hollowell (1990), Powell and Cosgrove (1992)]. While a purely quantitative output measure has been used (the number of full-time equivalent children enrolled) an attempt has been made to control for quality differences across day care centers, by means of variables reflecting care intensity and staff characteristics. However, the relationship between the "true" output – the increase in knowledge, skills and capability imparted to each child – and the quantity and quality components is not explicitly spelled out in these analyses.

Lately, a methodology for formulating measurement models for the output of educational institutions has been developed by Hanushek and Taylor (1990), and applied by Grosskopf et al. (1991, 1992). The strength of this approach lies in the explicit recognition of the fact that achievement test scores, which traditionally have been used to measure educational output, are to a large extent determined by circumstances exogenous to the producer (the school).[7] These exogenous influences are taken into account by regressing the test scores on previous test results and variables reflecting the socio-economic status of the student body, and using the estimated residuals of the regression equation as measures of output. While undoubtedly an improvement over old practices, this procedure still leaves quite a bit to be desired. For instance, no account is made of the effect on the student's capability to assimilate higher level education.

Irrespective of whether the measurement model is implicit or explicit, the common problem of the traditional approaches is the maintained hypothesis that the difference between the "true" output and the instrument by means of which it is modeled is equal to white noise. Only then is the use of output proxy variables a riskless undertaking, in the sense that it will not bias the conclusions drawn from the analysis. Being maintained, the white noise hypothesis cannot be tested. Quite often, it can be seriously questioned, however. Consider, for example, the almost universal practice of using patient-days as a measure of output in health care studies. Clearly, this measure may reflect next to nothing of the change in the present and future health status of patients, although this is something that ought to be included in any reasonable definition of health care output. Thus, there is no reason to believe that the number of patient-days mirrors the true output so closely so

---

[7] Of course, this insight is not new; it can be found, e.g., in Kiesling (1967).

that the two differ merely by a white noise term. Rather, the difference will most likely be systematic and perhaps of substantial magnitude.[8]

3. The alternative approach: imposing constraints on the production technology

The fact that the validity of the maintained hypothesis underlying the traditional approach often can be put into question makes it interesting to examine whether it is possible to do without quantitative output measures altogether. Obviously, when conditions are such that productivity and efficiency can be measured without quantitative information about the production result, then disregarding the by presumption defective output measure is the right thing to do. Furthermore, it reduces the part of the work associated with data collection and eliminates the need to formulate a measurement model. The problem, of course, is (i) to derive such conditions and (ii) to make sure that they are fulfilled.

The traditional approaches can give no hint about how to tackle (i) because they treat the output measurement problem in isolation, without taking the ultimate purpose, i.e. the measurement of productivity and efficiency, into account. In contrast, the analytical framework proposed here

[8] Strangely enough, these obvious shortcomings of the patient-days measure are rarely commented upon. One of the few exceptions is Grosskopf and Valdmanis (1987) who defend its use by claiming that their ambition is limited to modeling the production of an intermediate good, namely health services, rather than the final good, i.e. the change in (present and future) health. However, it does not seem very meaningful to consider the production of the intermediate good unless one can feel confident about knowing how it affects the production of the final good.

makes explicit use of the economic theory of production.[9] Specifically, it exploits the seminal work of Shephard (1953). Among other things, Shephard proved the following three properties for the cost function, which gives the minimum total cost of production as a function of the prices of the inputs employed and the volume of output produced. First, the cost function embodies all the essential information about the production technology. Thus, if the producer minimizes the cost of production, the technology that he/she employs can be characterized by means of the cost function. Second, the cost shares for input i can be obtained by differentiating the logarithm of the cost function with respect to the logarithm of the ith input price; this result is known as *Shephard's lemma.* Finally, for *homothetic* production technologies the cost function can be written as the product of two subfunctions: one function of output and one function of the input prices.

Homotheticity is the basic condition used in the thesis to enable productivity and inefficiency measurement without quantitative output data. Homothetic technologies can be characterized in two alternative ways. One characteristic is that the input cost shares are unaffected by the volume of output produced. This follows directly from the last two properties mentioned in the preceding paragraph. Homothetic technologies can also be described by the fact that the cost-minimizing production plan can be determined by means of the following two-step procedure. In the first step inputs are chosen such that the cost of producing one unit of output is minimized. The volume of output is then chosen such that profits are maximized, conditional on the input choice in the first step.

---

[9] To the best of my knowledge, the only previous attempt in this spirit is Hulten's (1984) study of the productivity in the public sector. However, Hulten's approach differs from the one taken here in that it is based on *household* production theory and presumes that the whole economy can be treated as a household. For a discussion of Hulten's model and its relation to this study, see Chapter II.

Chapter II and Chapter III are based on the first characterization. In these analyses it is assumed that there is no information at all about output. The derived models are primarily intended for studies of the production of public services. In Chapter IV, the second characterization is used. There it is assumed that, in addition to input data, there is information available about the value of output, but not about the output quantity or the output price. This situation is typical for many types of private services.

Of course, the homotheticity assumption is merely the starting point of the analyses in Chapters II, III, and IV. Further assumptions are necessary in order to enable conclusions about productivity and efficiency measurement. Both the assumptions and the conclusions will be commented upon in the next section. Before that, the possibility to ensure that the homotheticity requirement is fulfilled will be discussed.

In a strict sense, the only way to infer if a technology is homothetic or not is to model it as belonging to the more general class of non-homothetic technologies, which includes the class of homothetic technologies as a special case. One can then test if the assumption that the technology is homothetic amounts to imposing a binding restriction on the non-homothetic technology that one started out with. The only problem with this procedure is that it requires quantitative information about output. The reason is that in terms of the input cost shares, the difference between homothetic and non-homothetic technologies is that the cost shares corresponding to the latter are *not* independent of the volume of output. Some reflection shows that while it is not necessary to have perfect, error-free, output data in order to perform the test, the output proxy variables used (or some function of them) must have the property that they differ from the true output only by a white noise term. Now, from the discussion above it is clear that if this is the case then there is no reason not to use the output proxy variables. That is to say, the only situation in which it is possible to test if the homotheticity assumption is warranted is the situation in which there is no point in imposing this assumption.

This conclusion can be formulated in an alternative, considerably more interesting, way. To this end, assume that a production process is to be studied and that there is no *a priori* information about the production technology. Furthermore, assume that it will be possible to obtain output proxy variables but that there is no way of determining whether the difference between (a function of) these proxy variables and the true output will be merely white noise. The question then is: is the assumption that the technology is homothetic in some sense more restrictive than the assumption that the difference between (a function of) the output proxy variables and the true output is white noise? The answer to this question is that it cannot be answered. Thus, in this quite realistic situation it is impossible to say that the homotheticity assumption is too restrictive. The choice between this assumption and the use of proxy variables of unknown quality has to be determined by means of other criteria. As pointed out above, the homotheticity assumption has the advantage that it requires less data. Of course, if proxy variables are easily accessible at low cost then there is no need to make a choice; the two approaches can then be regarded as complementary.

Still, it might be argued that available experience speaks strongly against the homotheticity assumption because in studies where it has been tested it has mostly been decisively rejected. In the present context, the value of this empirical evidence is rather uncertain, however, since almost all of the tests come from studies of the manufacturing industry.[10] Analyses of

[10] Homotheticity has been tested in a few studies of service production, e.g. Spady and Friedlaender (1978) and Mester (1987), both of which reject it. However, the reliability of these tests is contingent on the assumption that the proxy variables used differ from the true output merely by white noise.

manufacturing are not of primary interest here because, as noted earlier, the output measurement problem is presumably much less severe in manufacturing than in the service sector.

More importantly, there is reason to expect the technologies employed in the production of services to differ from those utilized in the production of goods. If so, the fact that the homotheticity assumption is rejected in manufacturing does not necessarily mean that it is invalid with respect to the service sector. Indeed, it can be argued that the scope for automatization is more limited in the service industry than in the manufacturing industry and that this points in just the opposite direction. Because, compared to manufacturing, the possibilities to expand the productive capacity by increasing the capital stock (relative to the other factors of production) are mostly quite small; expansion often takes place by setting up new production units, similar to those already existing. Accordingly, *ceteris paribus*, the input cost shares may not change very much when the volume of production goes up (or down). Of course, this is just a conjecture; whether it is right or wrong is an empirical matter.

4. A brief summary of the three studies

At the cost of some repetition the three studies have been written in the form of independent articles that can be studied separately. It should be noted that the notation differs slightly between the chapters.

CHAPTER II: *On the econometric analysis of production when there are no output data*

Only input data are assumed to be available, in the form of quantities and prices. The properties of the input cost shares corresponding to a general homothetic cost function are investigated. Productivity measurement is made possible by allowing for (disembodied) technical change which, in accordance with common practice, is modeled as a function of time.[11] It is shown that if there are constant returns to scale, i.e. if the average cost is independent of the volume of output produced, then the growth in (total factor) productivity can be inferred merely by means of the system of input cost shares.[12] If there are non-constant returns to scale then the rate of productivity growth cannot be determined, but it will still be possible to infer whether it has increased or decreased from one period to another.

Two types of inefficiencies are considered and modeled parametrically: technical inefficiency and allocative inefficiency. Technical inefficiency means that the volume of output produced could have been accomplished by means of

---

[11] Disembodied technical change is that part of technical progress from which the producer can benefit effortlessly and costlessly. In contrast, embodied technical change can only be assimilated through the acquisition of new capital goods.

[12] Technologies exhibiting constant returns constitute a special case of homothetic technologies for which the average cost in general varies with the volume of output produced. Total factor productivity growth refers to the difference between the rate of change in output and a weighted average of the rates of change in all inputs.

smaller input quantities than those actually used. Allocative inefficiency refers to situations where the input proportions are inconsistent with cost minimization, given the relative input prices.

Modeling inefficiencies by means of the system of input cost shares is no trivial matter because the cost shares are derived under the assumption of cost minimization (cf. Section 3 above) while at the same time the existence of inefficiencies implies that costs cannot be minimized. The trick is to model the producer as minimizing cost subject to another technology and another set of input prices than those actually observed. This yields a generalized system of cost shares which contains the cost shares corresponding to cost minimization as a special case. Concerning allocative inefficiency, a previously used formulation is employed. Technical inefficiency is modeled in a new way, however. It is shown that the resulting generalized system of cost shares can be used to estimate the increases in total costs brought about by the inefficiencies, as well as their effects on input utilization.

CHAPTER III: *Identification of technical and allocative inefficiencies in the absence of output data*

This chapter is concerned with the econometric estimation of the model developed in Chapter II. To this end, a particular functional form has to be assumed for the cost function. The flexible and well-known translog cost function is chosen. The analysis is primarily motivated by the need to check that all parameters of the model are identified, in particular those intended to capture the presence of technical inefficiency. Since the model is highly non-linear, identification has to be established by means of a very general criterion.

The estimator considered is Full Information Maximum Likelihood, derived under the assumption that the disturbances in the cost share equations are joint normally distributed. Following Rothenberg (1971), identification is ascertained by proving that the information matrix has full rank.

CHAPTER IV: *An indirect approach to measuring productivity in private services*

In this chapter, the analysis in Chapter II is extended to the case where a measure of the value of output is available. Furthermore, imperfect competition is allowed for on the output market by modeling it as oligopolistic.[13] It is shown that if average cost is strictly increasing in the volume of output produced then the production process can be completely characterized, in spite of the presumed lack of quantitative output data.[14] This means that, in contrast to when only input data are available, it is possible to study scaling properties, i.e. the dependency of average costs on the volume of production. Accordingly, total factor productivity can be measured even in the case of non-constant returns to scale. This requires, however, that there is either perfect competition on the output market or, alternatively, that the markup that the sellers can charge over marginal costs is known to the researcher.

Since it is often difficult to obtain information about the markup the analysis proceeds to the case when the markup is unknown but constant. It is shown that the results continue to hold, conditional on the unknown markup.

Implementation of the results is illustrated by means of the translog cost function. It is shown that, in addition to yielding estimates of total factor productivity growth, the estimated parameters can also be used to compute a quantity index and a price index for the unknown output.

---

[13] Inefficiency measurement is not discussed in this chapter but the results on inefficiency obtained in Chapter II and Chapter III can be applied here, too.

[14] Although it is not mentioned in the article, the result holds also if average cost is strictly *de*creasing in the volume of output.

References

Berndt, E.R, B.H. Hall, R.E. Hall, and J.A. Hausman, 1974, Estimation and Inference in Nonlinear Structural Models, *Annals of Economic and Social Measurement* 3/4, pp. 653 – 665.

Bernstein, J.I., 1992, Information Spillovers, Margins, Scale and Scope: With an Application to Canadian Life Insurance, *Scandinavian Journal of Economics* 94 Supplement, pp. 95 – 105.

Eakin, B.K. and T.J. Kniesner, 1988, Estimating a Non-Minimum Cost Function for Hospitals, *Southern Journal of Economics* 54, pp. 583 – 597.

Grosskopf, S. and V. Valdmanis, 1987, Measuring Hospital Performance: A Non-parametric Approach, *Journal of Health Economics* 6, pp. 89 – 107.

Grosskopf, S., K. Hayes, L. Taylor, and W. Weber, 1991, Allocative Inefficiency in Education, Federal Reserve Bank of Dallas, Research Paper No. 9118.

Grosskopf, S., K. Hayes, L. Taylor, and W. Weber, 1992, Budget Constrained Frontier Measures of Fiscal Equality and Efficiency in Schooling, IUI Working Paper No. 335.

Gyapong, A.G. and K. Gyimah-Brempong, 1988, Factor Substitution, Price Elasticity of Demand and Returns to Scale in Police Production: Evidence from Michigan, *Southern Economic Journal* 54, pp. 863 – 878.

Hamermesh, D.S., 1986, The Demand for Labor in the Long Run, Ch. 8 in O.C. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. I, Amsterdam: North-Holland, pp. 429 – 471.

Hanushek, E.A. and L. Taylor, 1990, Alternative Assessments of the Performance of Schools, *Journal of Human Resources* 25, pp. 179 – 201.

Hulten, C.R., 1984, Productivity Change in State and Local Governments, *Review of Economics and Statistics* 66, pp. 256 – 266.

Jöreskog, K.G. and D. Sörbom, 1989, *LISREL 7 – A Guide to the Program and Applications*, 2nd ed., Chicago : SPSS Publications.

Kiesling, H.J., 1967, Measuring a Local Government Service : A Study of School Districts in New York State, *Review of Economics and Statistics* 49, pp. 346 – 367.

McKay, N.L., 1988, An Econometric Analysis of Costs and Scale Economies in the Nursing Home Industry, *Journal of Human Resources* 23, pp. 57 – 75.

Mester, L.J., 1987, A Multiproduct Cost Study of Savings and Loans, *Journal of Finance* 42, pp. 423 – 445.

Mukerjee, S., A.D. Witte, and S. Hollowell, 1990, Provision of Child Care : Cost Functions for Profit-Making and Not-for-Profit Day Care Centers, National Bureau of Economic Research Working Paper No. 3345.

Parsons, D.J., C.G. Gotlieb, and M. Denny, 1992, Productivity and Computers in Canadian Banking, forthcoming in *Journal of Productivity Analysis* 3.

Pauly, M.V., 1978, Medical Staff Characteristics and Hospital Costs, *Journal of Human Resources* 13, Supplement, pp. 77 – 111.

Powell, I. and J. Cosgrove, 1992, Quality and Cost in Early Childhood Education, *Journal of Human Resources* 27, pp. 472 – 484.

Rothenberg, T.J, 1971, Identification in Parametric Models, *Econometrica* 39, pp. 577 – 591.

Shephard, R.W., 1953, *Cost and Production Functions*, Princeton: Princeton University Press.

Spady, R.H., and A.F. Friedlaender, 1978, Hedonic Cost Functions for the Regulated Trucking Industry, *Bell Journal of Economics* 9, pp. 159 – 179.

Usher, D., 1980, ed., *The Measurement of Capital*, Chicago: University of Chicago Press for the National Bureau of Economic Research.

CHAPTER II

On the econometric analysis of production when there are no output data

(**with Bengt-Christer Ysander**)

1. Introduction

For many productive activities it is very difficult to define a relevant output measure – and often practically impossible to implement it, once defined. In particular, this is the case regarding the rapidly expanding service sector. Most services have several quality and quantity dimensions, some of which are largely unobservable. For instance, health care not only results in actual changes of patients' health status. It also helps prevent future, potential illnesses. Obviously, to quantify the latter effect is an almost hopeless task.

The most severe output measurement problems are probably encountered in the public sector. In the national accounts system, this has lead to the convention that the value of a public service is set equal to the value of the resources used to produce it. Volume measures are obtained by weighing the inputs by constant, rather than current, prices. Accordingly, the volume of a particular service, q say, in year t is defined as

$$q_t \equiv x_{1t} + x_{2t} + \cdots + x_{nt}$$

where x_i is the amount used of input i, valued at constant prices.[1] This accounting practice implies several strong assumptions about the productive performance of the public sector, some of which do not seem to have been

[1] In some countries, e.g. the U.S., only labor input is considered.

generally recognized.

Firstly, it implies that the growth in total factor productivity, defined as the difference between the growth in real output and the growth of the cost share weighted inputs according to, e.g., Jorgenson and Griliches (1967), will always be equal to zero. Apart from almost certainly yielding an incorrect measure of the productivity development in the public sector itself, this will also bias calculations of aggregate growth (e.g. GDP), as soon as the size of the public sector changes.

Secondly, it means that the production in the public sector is assumed to be efficient in the sense that there is no slack in the utilization of the various factors of production. If such slack were to exist it would be possible to reduce the usage of some of the inputs without reducing output but, according to the chosen method of measurement, any such reduction would decrease the level of output. This is in contrast not only with widely held beliefs but also with theoretical considerations predicting lower efficiency in the public than in the private sector.

Thirdly, it can readily be seen that a proportionate increase in all inputs will increase q by the same proportion, implying that constant returns to scale are assumed. In view of the fact that diminishing average costs, i.e. increasing returns, is an important motivation for public production, this is somewhat unfortunate.

Finally, the additive formulation amounts to assuming that all inputs are perfect substitutes. However, given this technological property a cost-minimizing producer would of course only use the cheapest input. Hence, it must be implicitly assumed that public producers ignore the effects of relative prices on total costs. Although this seems to agree quite well with common opinion it would be preferable to regard it as an hypothesis to be tested rather than as a *maintained* hypothesis.

In this theoretical paper we show that given (time series) input data, three of these four issues, namely productivity growth, efficiency in production, and sensitivity to changes in relative input prices, are amenable to econometric analysis *even in the absence* of output measures. There is thus no need to arbitrarily determine them *a priori*. Moreover, although we cannot explicitly study properties concerning returns to scale, our approach allows for the possibility of variable returns.[2]

Our results are completely general in the sense that they can be applied to *any* production activity, i.e. not only to those in the public sector. Within the private sector, the production of banking services is an example of an interesting object of study. In the national accounts, value measures of the output in the banking sector are obtained by adding the bank's service charges and the net proceeds from their lending operations. To construct volume measures of output, various *ad hoc* assumptions are made. In the Swedish national accounts, e.g., it is assumed that the banking industry every year experiences a 2% increase in average labor productivity.[3] Similar procedures are employed in other countries, too. Our approach makes it possible to investigate the empirical validity of such assumptions.

In contrast to the method that we are going to propose, analyses of production activities for which there are no reliable output measures traditionally have employed proxy variables, intended to mirror the unknown output.

---

[2] Of course, this is not to say that one should not make efforts to search for appropriate and operational output measures. Such data enable more elaborate analyses and the measurement of returns to scale. The purpose of this paper is merely to show how much information one can obtain about the production process in situations where either no output data are available or it is impossible to assess the quality of existing data.

[3] As far as we know, no empirical support exists for the particular choice of 2%. It is interesting to note that in an attempt to measure average labor productivity in American banks over the period 1927-1979, Rhoades and White (1984) could not find any indication of growth in average labor productivity since the mid 1950's.

Attempts have also been made to take several dimensions of output into account simultaneously, either by aggregating several proxies into an output indicator index or by modeling production as multiple-output processes. Still, studies of this kind can always be criticized for failing to account for such basically unobservable output dimensions as the one exemplified in the first paragraph above. Since, by their very nature, such characteristics cannot be explicitly incorporated into the analysis the only way to escape this criticism is to find some method of avoiding the measurement of output altogether, as we do in this paper. To our knowledge, the only previous attempt in this vein is Hulten's (1984) study of productivity changes in the public sector.

Inspired by household production theory, Hulten models the whole economy as a "household", maximizing a utility function in an aggregate private sector good, directly available for consumption, and an aggregate public sector commodity, which is produced by the community for internal consumption. The production process yielding the public sector commodity is assumed to exhibit constant returns to scale. Productivity changes are further presumed to be Hicks-neutral and are modeled by an exponential time trend. Duality theory can then be used to express the price of the public sector commodity in terms of the prices of the factors of production and a time index. Hulten demonstrates that this result in turn makes the ratio of the "household" budget shares for the private and public sector outputs a function of the price of the private sector good, the factor prices and the time index. By means of this equation the rate of public sector productivity growth can be estimated without an explicit measure of the public sector output.

In addition to the rather restrictive assumptions about the production technology a serious problem with Hulten's approach is the maintained, and therefore untestable, hypothesis that the household/community analogue is indeed valid, which is far from obvious. Our method is based only on standard neoclassical production theory and, hence, can be applied to the public sector without any such assumption. Moreover, in contrast to Hulten, we do not have to presume the availability of any other information than input data for the particular production process studied.

Given only input data, production or profit functions are infeasible as instruments of analysis, since in studies based on these the level of output is endogenously determined. This leaves a cost function analysis, where the output level is treated as predetermined, as the only practical alternative.[4]

Output predeterminacy alone will not make it possible to analyze the production process by means of input data only. Both the cost function and the input demands which can be derived from it will always be dependent on the level of output. However, if the production technology is homothetic, i.e if the proportions in which the factors of production are employed are unaffected by the scale of operation, then the shares of the various inputs in total cost will be independent of the output level. The input cost shares will thus be the endogenous variables in our analysis.

The property that the input cost shares of a homothetic technology are invariant to the level of output has long been recognized in the econometrics literature. The extent to which these cost shares can yield information about the production process has not been thoroughly investigated, however. In this paper we perform such an investigation, based on a homothetic cost function formulated in general terms.

The paper unfolds as follows. In Section 2, some well known implications of homotheticity are briefly stated. Estimation of the effects of non-

[4] The treatment of output as a predetermined variable does not necessarily imply that the output level is exogenous to the producer. It can be justified even if the output decision is taken by the producer himself, provided that the problem of minimizing unit costs can be separated from the problem of choosing the level of output. This independence condition is fulfilled by the homothetic technologies that we will consider here.

neutral technical change on input requirements, total costs, and on total factor productivity is taken up in Section 3. In Section 4 we consider the fact that the theoretical derivation of the input cost shares assumes that production costs are minimized. We demonstrate how the dual representation can be generalized to allow for the existence of technical inefficiency (overutilization of inputs) and allocative inefficiency (inoptimal factor proportions), implying higher than minimum costs. Moreover, we show that this generalized system of cost shares can be used to estimate the increases in total costs brought about by the inefficiencies, as well as their effects on input utilization. As far as we know, the fact that this is possible even when there are no data on output has not been demonstrated earlier. Formulas for comparing price and substitution elasticities, and the estimated effects of technical change, when there are inefficiencies in production with the corresponding measures under cost minimization are also given. Concluding comments are given in Section 5.

2. The meaning of homotheticity

We assume that information is available about the quantities used of the different factors of production and their respective prices, but that there are no data on output. To simplify the discussion, we will in this section disregard technical change and potential inefficiencies in production. For the time being we thus assume a static technology and cost-minimizing producers.

Let the minimum cost function be $C = C(y,\boldsymbol{w})$, where y is the unknown output ($y > 0$) and $\boldsymbol{w}$ denotes the vector of (strictly positive) input prices, $\boldsymbol{w} = (w_1,...,w_n)$.[5] The cost function must fulfill certain regularity conditions which, e.g., can be formulated as in Diewert (1971, pp. 489-90). To be regular, $C = C(y,\boldsymbol{w})$ should be non-decreasing in both y and $\boldsymbol{w}$, and be linearly homogeneous and concave in $\boldsymbol{w}$. If these conditions are all satisfied, the cost function will describe all economically relevant aspects of the production technology. In addition, it will be assumed here that $C = C(y,\boldsymbol{w})$ is twice differentiable with respect to each of its arguments.

As shown by Shephard (1953, pp. 45-47), if the production technology is homothetic then the cost function is separable y and $\boldsymbol{w}$, according to

$$C = C(y,\boldsymbol{w}) = f(y) \cdot g(\boldsymbol{w}) \,, \tag{1}$$

where f is a monotonically increasing function of y. By means of restrictions on $f(y)$ the homothetic technology can be specialized into a homogeneous technology. In particular, linear homogeneity, i.e. constant returns to scale, requires that f be equal to the identity function. It can be shown that given an appropriate definition of $f(y)$, the function $g(\boldsymbol{w})$ equals the cost of producing one unit of output, i.e. $C(1,\boldsymbol{w})$.

It is the function $g(\boldsymbol{w})$ that we will be interested in. As separate identification of $f(y)$ and $g(\boldsymbol{w})$ requires some kind of normalizing restriction we will assume that $g(1) = 1$.[6]

---

[5] As pointed out to us by Rolf Färe, the results in the following are valid not only for single output technologies, but for multiple output technologies as well. We could thus replace the scalar y by an m vector $\boldsymbol{y} = (y_1,...,y_m)$ of outputs. However, since we have postulated the non-existence of output data the distinction between these two cases becomes rather subtle so, for simplicity, we treat output as a scalar quantity.

[6] This is a convenient normalization in applied work where the elements of $\boldsymbol{w}$ are often price indices rather than (absolute) price levels. The normalization will then ensure that the base-year for the unit cost index, i.e. $g(\boldsymbol{w}) = C(1,\boldsymbol{w})$, is the same as the base year for the input price indices.

By *Shephard's lemma* the producer's input demands are given by:

$$x_i = x_i(y,\boldsymbol{w}) = f(y)\cdot\frac{\partial g(\boldsymbol{w})}{\partial w_i}, \qquad i = 1,\dots,n. \tag{2}$$

In accordance with (1) and (2) the input cost shares can be written

$$s_i = s_i(\boldsymbol{w}) = \frac{w_i\cdot\frac{\partial g(\boldsymbol{w})}{\partial w_i}}{g(\boldsymbol{w})}, \qquad i = 1,\dots,n. \tag{3}$$

In contrast to the input demands, the cost shares are independent of the level of output, y. It is thus possible to estimate the system (3), and hence the function $g(\boldsymbol{w})$, without having to take the level of output into account. This obviously solves our main problem, i.e. that of eliminating the unknown entity y from the analysis.

It is clear, however, that the system (3) cannot provide a full description of the production technology as it does not yield information about the function $f(y)$.[7] In contrast, the system (2) of input demands contains all the information available in the original cost function, since the input demands multiplied by the factor prices add up to $C(y,\boldsymbol{w})$, by Euler's theorem. This difference in informational content between the two systems is explained by the fact that whereas the system (2) is of full rank (i.e. n) the rank of the system (3) of cost shares is only n-1, which can easily be seen by noting that both sides of (3) sum identically to one. As a consequence, one of the share equations must be dropped when the system is estimated.[8] This, in turn, implies that if the functional form chosen for C is *flexible* symmetry has to be

---

[7] Given an estimate $g^*(\boldsymbol{w})$ of $g(\boldsymbol{w})$ an estimate of $f(y)$ can be obtained by means of the ratio $\boldsymbol{w}'\boldsymbol{x}/g^*(\boldsymbol{w})$ where $\boldsymbol{w}'\boldsymbol{x}$ is observed total cost. The form of the function f and the value of y cannot be inferred, however, except in the special case when there are constant returns to scale, implying that $f(y) = y$.

[8] If the estimation method is maximum likelihood and the stochastic disturbance terms are additively appended to the equations (3) the estimation results will be invariant to the choice of the left out equation, cf. Barten (1969).

imposed *a priori* to ascertain identification of the function $g(\boldsymbol{w})$.

The information loss incurred by studying the system of input cost shares only concerns the scaling properties of the technology, however. Factor substitution and the price responsiveness of input demands can still be studied. Using the results of Uzawa (1962), the Allen partial elasticities of substitution [Allen (1959)] can be expressed in terms of the input cost shares and the factor prices according to

$$\sigma_{ij} = \frac{C \cdot \dfrac{\partial^2 C}{\partial w_i \partial w_j}}{\dfrac{\partial C}{\partial w_i} \dfrac{\partial C}{\partial w_j}} = \left[s_i s_j + w_j \cdot \frac{\partial s_i}{\partial w_j} \right] \cdot \frac{1}{s_i s_j}, \tag{4}$$

while the price elasticities can be calculated as

$$\eta_{ij} \equiv \frac{\partial x_i}{\partial w_j} \frac{w_j}{x_i} = s_j \sigma_{ij}\ . \tag{5}$$

The homotheticity assumption can of course be questioned. In a static environment – i.e. in the absence of technical change – it implies that the cost-minimizing input mix is determined by relative input prices only, which is often a restrictive assumption. As a consequence, with constant relative prices the expansion path will be linear [cf. Färe (1974)]. This may not be consistent with the often noted tendency to increase the capital intensity at larger scales of operation.[9]

Homotheticity has been decisively rejected in many applied production studies. It can be argued, however, that the homotheticity assumption is easier to defend in the context of service production than in the production of

9 Changing the scale of operation generally takes some time, during which relative input prices may change, too. Thus, an observed increase in the capital intensity during an output expansion need not necessarily be inconsistent with homotheticity.

goods, because services are more difficult to routinize, making the scope for automatization more limited. Compared to for example manufacturing, the possibilities to expand the productive capacity by increasing the capital stock relative to the other inputs are often quite small; expansion often takes place by setting up additional production units, similar to those already existing. Accordingly, *ceteris paribus*, the input cost shares may not change very much when the volume of production is increased (or decreased). Although this argument should be used with caution the homotheticity assumption appears to be more applicable where it is more needed, i.e. in service production where no reliable output measures are available. In the case of government services, homotheticity may, moreover, reflect centralized decision making which tends to treat establishments of different size – e.g. schools – all alike.[10]

3. Technical change and total factor productivity

In accordance with common practice, we assume that technical change (of a disembodied nature) can be modeled by means of a time index, t.[11] The cost function can then be formulated according to

$$C = C(y,\mathbf{w},t) = f(y,t) \cdot g(\mathbf{w},t) . \tag{6}$$

This formulation encompasses three types of technical change: non-neutral

[10] It should be noted that there is no obvious conflict between centralized decision making and cost minimization. If the central decisions take the form of requirements on the input mix, conditioned upon a given set of factor prices, they may have precisely the effect of imposing a homotheticity constraint on the production possibilities facing the local producers. As long as the central decrees are optimally adjusted to changes in the relative input prices, costs will be minimized, albeit subject to a homotheticity restriction.

[11] See, for example, Binswanger (1974), Berndt and Khaled (1979), Nadiri and Schankerman (1980) and Parsons, Gotlieb and Denny (1992).

technical change, neutral and scale-independent technical change and, finally, neutral technical change which affects the scaling properties of the technology.

Technical change is said to be non-neutral or (Hicks-)neutral depending on whether it affects or does not affect the marginal rates of substitution between inputs, respectively. It will be shown that the effects of non-neutral technical change and neutral and scale-independent technical change an be estimated by means of input data only. These two forms of technical change can be captured by the function $g(\cdot)$. However, like scaling properties in general, changes in returns to scale induced by technical change cannot be measured without output data. For this reason we will consider the following restricted form of (6)

$$C = C(y,\boldsymbol{w},t) = f(y)\cdot g(\boldsymbol{w},t)\ . \tag{6'}$$

For empirical purposes the substitution of (6') for (6) is probably of little importance. In fact, the restricted form (6') is almost invariably used in applied studies, because of the practical difficulties encountered in attempts to (simultaneously) obtain estimates of all three forms of technical change.[12] It is fair to say, we believe, that compared to the homotheticity assumption, the assumption that technical change does not affect returns to scale is very weak. Given (6') the system of input cost shares becomes

$$s_i = s_i(\boldsymbol{w},t) = \frac{w_i \cdot \frac{\partial g(\boldsymbol{w},t)}{\partial w_i}}{g(\boldsymbol{w},\, t)}\ . \tag{7}$$

We begin by considering the effect of technical change on the cost shares, the input demands and on total costs. We then use the connection between

[12] None of the four studies mentioned in the previous footnote report succesful attempts to estimate the effects of all the three forms of technical change. Binswanger (1974) and Berndt and Khaled (1979) impose the constraint (6') a priori, while Nadiri and Schankerman (1980) and Parsons, Gotlieb and Denny (1992) start with (6) but fail to obtain significant estimates of the effects of technical change on the returns to scale.

technical change and total factor productivity to investigate what conclusions can be drawn about the rate of total factor productivity growth.

3.1 Effects on input cost shares, input demands, and total costs

By including the time index, the input cost shares are allowed to shift over time not only in response to changes in relative factor prices but also because of exogenously determined technological developments. In the following, we will use the letter τ to denote a relative time derivative. Accordingly, the relative effects of technical change on the cost shares – i.e. the Binswanger (1974) measures of the biases in technical change – will be written

$$\tau_{s_i} \equiv \frac{\partial s_i(\boldsymbol{w},t)}{\partial t} \frac{1}{s_i(\boldsymbol{w},t)}, \qquad i = 1,\dots,n. \tag{8}$$

If $\tau_{s_i} < 0$ technical change is characterized as relatively factor i-saving and if $\tau_{s_i} > 0$ it is said to be relatively factor i-using. In the presence of non-neutral technical change $\tau_{s_i} \neq 0$ for at least one $i = 1,\dots,n,$ thereby indicating that that technical affects the relative development of the input cost shares over time. If technical change is neutral then $\tau_{s_i} = 0 \;\; \forall\, i.$ [13]

If technical change is Hicks-neutral the partial derivatives $\partial s_i(\boldsymbol{w},t)/\partial t$ will be identically zero for all i and the system (7) will degenerate to the static technology system (3). The cost shares thus cannot be used to test the hypothesis that the technology has not been undergoing any form of technical change – if (7) is not found to be statistically superior to the static system (3) the technology might still have been subject to neutral technical change.

It should be noted though that the fact that the cost shares are invariant

[13] It has been shown by Blackorby, Lovell, and Thursby (1976) that for homothetic technologies the definitions of neutral and non-neutral technical change in terms of the marginal rates of substition and in terms of the effects on the input cost shares are equivalent.

to neutral technical change does *not* imply that it is useless to try to capture input neutral effects of technical change by means of these same cost shares. It is possible to construct models of technical change which have the properties that they (i) affect the input cost shares, i.e. are non-neutral and (ii) can be specialized to the Hicks-neutral case *and* which (iii) can be completely determined by the cost shares.[14] As an example, consider the following Cobb-Douglas cost function:

$$C(y,\boldsymbol{w},t) = y\cdot w_1^{\alpha} w_2^{1-\alpha}\cdot \exp\left[\beta t\cdot (w_1/w_2)^{\gamma}\right] .$$

For $\gamma = 0$ technical change is neutral. The input cost shares are

$$s_1 = \alpha + \beta\gamma\cdot tw_1(w_1/w_2)^{\gamma} \quad s_2 = (1-\alpha) - \beta\gamma\cdot tw_1(w_1/w_2)^{\gamma} .$$

By estimating one of the share equations we can obtain measures of both the neutral and the non-neutral components of technical change, provided that both are non-zero.[15]

The effects of technical change on input demands can also be estimated, in spite of the fact that the input demands are dependent upon the level of output. Since

$$x_i = x_i(y,\boldsymbol{w},t) = f(y)\cdot \frac{\partial g(\boldsymbol{w},t)}{\partial w_i}, \qquad i = 1,\ldots,n , \tag{9}$$

the rate of change in the demand for input x_i can be expressed in terms of only the input prices and the time index according to

$$\tau_{x_i} \equiv \frac{\partial x_i}{\partial t}\frac{1}{x_i} = \frac{\partial^2 g(\boldsymbol{w},t)}{\partial w_i \partial t}\cdot \left[\frac{\partial g(\boldsymbol{w},t)}{\partial w_i}\right]^{-1}, \quad i = 1,\ldots,n. \tag{10}$$

---

[14] We are grateful to Jacques Mairesse and Ishaq Nadiri for inspiring us to consider this issue.

[15] That the parameters are identified follows from the fact that the partial derivatives of s_1 with respect to α, β and γ are linearly independent.

Finally, it can be shown that the effect of technical change on total costs is given by the effects on the input demands weighted by the corresponding cost shares, i.e.

$$\tau_C \equiv \frac{\partial C(y,\boldsymbol{w},t)}{\partial t} \frac{1}{C(y,\boldsymbol{w},t)} = \sum_{i=1}^{n} s_i \cdot \tau_{x_i} \,. \qquad (11)$$

[16]

3.2 Total factor productivity

The effects of technical change on the rate of total factor productivity can be analyzed by means of a general duality result derived by Ohta (1974). Let $\psi(\boldsymbol{x},t)$ denote the production function to which the cost function (6') is dual. The *primal* rate of total factor productivity can then be defined according to

$$\tau_\psi \equiv \frac{\partial \psi(\boldsymbol{x},t)}{\partial t} \frac{1}{\psi(\boldsymbol{x},t)} \,.$$

What Ohta has shown is that the following dual relationship holds

$$\tau_\psi = (-\tau_C) \cdot (e_{Cy})^{-1} \qquad (12)$$

where τ_C is given by (11) and

$$e_{Cy} \equiv \frac{\partial C(y,\boldsymbol{w},t)}{\partial y} \frac{y}{C(y,\boldsymbol{w},t)} \,. \qquad (13)$$

The first factor in (12), the negative of the rate of change in total cost, is the dual representation of technical change. The second factor, the inverse of the elasticity of total cost with respect to output, is the dual form of the rate of return to scale. Returns to scale are increasing if $e_{Cy} < 1$, constant if $e_{Cy} = 1$, and decreasing if $e_{Cy} > 1$. It can be shown that for a homothetic

[16] We note in passing that

$$\tau_{s_i} = \tau_{x_i} - \tau_C \qquad i = 1,\ldots,n.$$

Hence, technical change may be input i-using even if it has the effect of diminishing the use of all n inputs.

technology e_{Cy} will always be strictly positive, see e.g. Førsund (1975).

Using (11) and applying (13) to (6') we get

$$\tau_{\psi} = \left[-\sum_{i=1}^{n} s_i \tau_{x_i}\right] \times \left[\frac{y \cdot f'(y)}{f(y)}\right]^{-1}. \tag{14}$$

Because of the occurrence of y in the last factor of (14), it is obvious that, in general, the system of input cost shares does not provide all the information needed to calculate an estimate of the rate of total factor productivity. However, since we know that e_{Cy} will be strictly positive the sign of (14) will be equal to the sign of the first factor on the right hand side, i.e. the dual rate of technical change. Accordingly, the question of whether total factor productivity is increasing or decreasing can always be answered by means of the first factor in (14), which can be obtained from the estimation of the system of cost shares.

If, furthermore, the technology is homogeneous then the rate of return to scale will be independent of the level of y and so the last factor in (14) will be equal to a constant, instead of being a function of y. In that case it will be possible to construct an index of total factor productivity growth because τ_{ψ} is determined up to a constant of proportionality. If, finally, the technology is linearly homogeneous, i.e. characterized by constant returns to scale, then e_{Cy} will be equal to unity and the negative of τ_C will be identical with the rate of change in total factor productivity.

The conclusions that can be drawn about productivity growth when only input data are available will thus depend on how restrictive assumptions we are willing to make concerning returns to scale. Thus, while the homotheticity assumption always allows us to determine the sign of the productivity growth rate we need to assume constant returns to scale to be able to obtain a complete characterization of the growth in total factor productivity.

4. Deviations from cost minimization

By definition, $C(y,\mathbf{w},t)$ denotes the smallest total cost attainable in time period t for input vectors yielding at least the output y. If costs are not minimized, estimation of the system of cost shares may yield biased estimates of price and substitution elasticities and of the effects of technical change on the production process. These considerations are of particular importance concerning public sector applications, as there are theoretical arguments for questioning cost minimization as the primary objective of public producers.[17]

Deviations from minimum costs – which, of course, must always be positive – are commonly taken to arise because of inefficient producer behavior, but there may be other reasons as well.[18] We will not try to discriminate between different sources of inefficiency, however. Following the literature in this field, we will be content with merely examining in what ways the *existence* of inefficiencies can be modeled and, secondly, how their effects on total costs and input demands can be estimated. Also in line with the literature, we will henceforth sometimes speak of the *degree of efficiency* instead of inefficiency. The degree of efficiency should be interpreted here as a (truncated) fractional measure such that a degree of efficiency in the open interval]0,1[implies a certain amount of inefficiency, whereas a degree of efficiency equal to one means (fully) efficient.

We have chosen to model inefficiency parametrically, which makes it possible to implement our results with a wide variety of flexible functional forms. Two types of inefficiency are considered: technical inefficiency,

---

[17] These arguments are summarized in Byrnes, Grosskopf and Hayes (1986).

[18] Another cause may be regulatory constraints, see for example Atkinson and Halvorsen (1984, 1986). Also, if the exogenously given demand is highly variable it may be impossible to avoid some slack in off-peak periods in order to be able to cope with the peaks, cf. Fuss and McFadden (1978).

concerning overutilization of inputs, and allocative or price inefficiency, referring to situations where the factor proportions are inconsistent with cost minimization, given the relative input prices.

Among the various concepts of input efficiency we are thus disregarding scale inefficiency.[19] The reason is, of course, that the presumed lack of output data makes it impossible to analyze scaling properties. However, the invariance of the cost shares with respect to the scale of production means that scale inefficiency will not introduce any bias in the conclusions that we actually are able draw by means of the system of input cost shares. It should also be noted that from the perspective of an individual producer, scale efficiency is not a well defined concept in the sense that it is not always consistent with cost minimization. The potential inconsistency arises when the level of output is exogenously given to the producer and the optimal scale, y^* say, i.e. the scale for which the dual scale elasticity $(e_{Cy})^{-1}$ is equal to unity, is greater than the exogenously given level of output, y. In that case an adjustment towards scale efficiency will increase total costs, since $C(y,\boldsymbol{w},t)$ is non-decreasing in y; cf. Section 2. In the following we will take the perspective of the individual firm. Thus, we will consider a producer which is both technically and allocatively efficient to be overall efficient and identify minimum total costs with the total costs incurred in the context of the so defined overall efficiency.

[19] For discussions of the various concepts of productive efficiency, see Førsund and Hjalmarsson (1974, 1979) and Färe, Grosskopf, and Lovell (1985). In addition to technical, allocative, and scale efficiency Färe te al. consider yet another input efficiency concept, namely that of (absence of) congestion. However, congestion can only arise when the technology is characterized by weak disposability of inputs (WDI), implying that an increase in the utilization of some input(s) may in some cases decrease the amount of output. Since free disposability of inputs (FDI) – increases in input can never decrease output – is one of the regularity conditions which have to be fulfilled to ascertain a dual representation of a production technology [cf. Diewert (1971)], WDI technologies, and thus congestion, are of no interest in our context.

We will begin by discussing how the system of cost-minimizing shares can be generalized to take allocative inefficiency into account, given that the production process is technically efficient. We then show that the resulting cost shares are invariant with respect to *radial technical inefficiency*, the technical inefficiency counterpart to neutral technical change. Next, we demonstrate that it is possible to formulate a cost share system which captures the *combined* effect of both allocative and technical inefficiency, but which cannot unambiguously distinguish between the two. We establish, however, that under weak conditions a well defined decomposition of the combined effect into radial technical inefficiency and allocative inefficiency can be obtained by means of the first specification, i.e. the one allowing explicitly for allocative and implicitly for radial technical inefficiency. Finally, we show how the computations of the price and substitution elasticities are affected by deviations from cost minimization and, similarly, how the various measures of technical change should be calculated.

4.1 Allocative inefficiency

As is well known, the first order conditions for cost minimization require that the inputs be chosen such that the ratio of their marginal productivity values, or shadow prices, be equal to the ratio of their (actual) prices. Since the marginal rates of technical substitution are equal to the corresponding ratios of marginal productivity values, this requirement can equivalently be expressed as requiring equality between

$$\frac{\partial\psi(\hat{\boldsymbol{x}},t)/\partial x_i}{\partial\psi(\hat{\boldsymbol{x}},t)/\partial x_j} \equiv \frac{w_i^\diamond}{w_j^\diamond} \qquad \text{and} \qquad \frac{w_i}{w_j}$$

for all $i \neq j$, where, as before, $\psi(\cdot)$ denotes the production function to which the cost function $C(\cdot)$ is dual, $\hat{\boldsymbol{x}} = (\hat{x}_1,..., \hat{x}_n)$ is the (hypothetical) point at which the marginal products are evaluated and $w_i^{\diamond}$ is the shadow price of input i. Using $w_n^{\diamond}$ and w_n to normalize, the production process can be defined as allocatively efficient if $w_i^{\diamond}/w_n^{\diamond} = w_i/w_n$ for $i = 1,\dots,n\text{-}1$.

A simple, yet quite powerful, specification by means of which deviations from allocative efficiency can be studied is the following one, originally proposed by Lau and Yotopoulos (1971) and introduced in a dual, time series context by Toda (1976, 1977).[20] The shadow prices are assumed to be proportional to the factor prices actually observed, according to

$$w_i^{\diamond} = \lambda_i^{\diamond} w_i\,, \qquad \lambda_i > 0\,, \qquad i = 1,\dots,n\,, \tag{15}$$

where $\lambda_i^{\diamond}$ is an input-specific proportionality constant to be estimated.[21] (The reason why we have attached the $\diamond$ to the proportionality constant will become clear in Section 4.2.) Since the first order conditions only concern relative prices the following normalization rule can be imposed without loss of generality

$$\lambda_n^{\diamond} = 1\,, \tag{16}$$

cf., for example. Atkinson and Halvorsen (1984).

The *realized* cost shares – as opposed to the cost minimizing shares – can be derived as follows. The producer's choice of input levels can be regarded as the result of minimizing total shadow costs, $\Sigma_{i=1}^{n} w_i^{\diamond} x_i$, with respect to the x_i, $i = 1,\dots n$. Using (6') and (15) – (16) the *minimum total shadow costs* can be expressed in terms of the actually observed input prices

---

[20] Toda considered the two input case. The generalization to the n input case which we use in the following is due to Atkinson and Halvorsen (1984).

[21] It is of course possible to model deviations from allocative efficiency in other ways, too. Eakin and Kniesner (1988) have proposed an additive, rather multiplicative, relation between the shadow prices and the actual input prices. The specification (15) is, however, by far, the one most commonly used.

according to

$$C(y,\boldsymbol{w}^{\diamond},t) = f(y) \cdot g(\boldsymbol{w}^{\diamond},t)\,, \tag{17}$$

where

$$\boldsymbol{w}^{\diamond} = (w_1^{\diamond},\ldots, w_n^{\diamond})' = (\lambda_1^{\diamond} w_1,\ldots, \lambda_{n-1}^{\diamond} w_{n-1}, w_n)'\,.$$

It is obvious that $C(y,\boldsymbol{w}^{\diamond},t)$ fulfills the regularity conditions cited in Section 2 when the input price vector is taken to be $\boldsymbol{w}^{\diamond}$. Hence, $C(y,\boldsymbol{w}^{\diamond},t)$ is a proper dual representation of some underlying production technology, given the price vector $\boldsymbol{w}^{\diamond}$. Application of Shephard's lemma to (17) yields the input levels which minimize total shadow costs, the $\hat{x}_i$'s, as

$$\hat{x}_i = \frac{\partial C(y,\boldsymbol{w}^{\diamond},t)}{\partial(\lambda_i^{\diamond} w_i)} = f(y) \cdot \frac{\partial g(\boldsymbol{w}^{\diamond},t)}{\partial(\lambda_i^{\diamond} w_i)}\,, \quad i = 1,\ldots,n\,. \tag{18}$$

Using (18), the *realized total cost*, C^a, can be written

$$C^a \equiv \sum_{i=1}^{n} w_i \hat{x}_i = f(y) \cdot \sum_{i=1}^{n} w_i \frac{\partial g(\boldsymbol{w}^{\diamond},t)}{\partial(\lambda_i^{\diamond} w_i)} \tag{19}$$

which differs from $C(y,\boldsymbol{w}^{\diamond},t)$ because the partial derivatives are weighted by the actual input prices, the w_i's, rather than by the shadow prices, the $w_i^{\diamond}$'s. It can be shown that $C^a \geq C$, at least as long as $g(\boldsymbol{w},t)$ is concave in $\boldsymbol{w}$.[22] If $\lambda_i^{\diamond} = 1$ for all i then C^a, $C(y,\boldsymbol{w}^{\diamond},t)$, and $C(y,\boldsymbol{w},t)$ are all identically equal.

The system of cost shares to be estimated will thus be

$$s_i^a \equiv \frac{w_i \cdot \hat{x}_i}{C^a} = \frac{w_i \cdot \dfrac{\partial g(\boldsymbol{w}^{\diamond},t)}{\partial(\lambda_i^{\diamond} w_i)}}{\sum_{k=1}^{n} w_k \cdot \dfrac{\partial g(\boldsymbol{w}^{\diamond},t)}{\partial(\lambda_i^{\diamond} w_i)}}\,, \quad i = 1,\ldots,n. \tag{20}$$

[22] Toda considered the two input case. The generalization to the n input case which we use in the following is due to Atkinson and Halvorsen (1984).

In the estimation, the positivity constraints on the $\lambda_i^\diamond$'s [cf. (15)] constitute a potential estimation problem. To ascertain that the $\lambda_i^\diamond$'s stay positive they can be defined in terms of a transformation function, according to $\lambda_i^\diamond \equiv \varphi(\beta_i)$ where β_i is an unrestricted parameter and φ a function whose image is equal to the set of positive real numbers. For instance, φ might be an exponential function as suggested by Lau (1978).

Testing allocative efficiency means testing the hypothesis that all the $\lambda_i^\diamond$'s are equal to unity, in which case (20) is identically equal to the system of cost minimizing shares. Fig. 1 illustrates the test in the two input case.

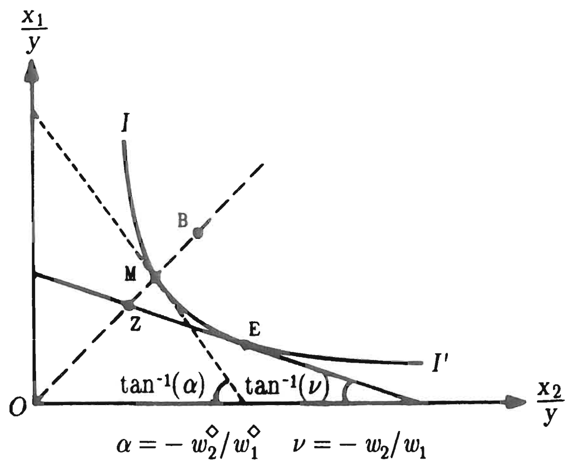

Figure 1. Farrell measures of allocative, technical, and overall efficiency

To be capable of illustrating both allocative and technical inefficiency the diagram is drawn in the space of input/output-coefficients. Thus, all points lying on or to the northeast of the isoquant II' correspond to the same volume of output. The isocost shown by a solid line corresponds to the factor prices actually observed, i.e. w_1 and w_2. Since we are here assuming that the producer is technically efficient, production must be taking place somewhere along the isoquant II' . The producer will minimize costs by operating at the point *E*. Assume, however, that production is actually taking place at the point *M*. With the input prices at the observed levels this point is obviously not allocatively efficient. However, *M* would have been an allocatively efficient location had the isocost been given not by the solid but by the dotted line. The slope, α, of this latter isocost equals the ratio of the shadow prices since, given that production occurs at *M*, this is the relative price corresponding to cost minimization. The hypothesis to be tested is thus whether the slope of the hypothetical isocost, α, is significantly different from ν, the slope of the actual isocost. In the two input case this simply means testing if $\lambda_1^\diamond = 1$ since, in accordance with (16), $\lambda_2^\diamond = 1$ a priori.

Farrell (1957) proposed a scalar measure of the degree of price efficiency. In terms of Fig. 1, Farrell's measure of allocative efficiency *(AE)* is defined as

$$AE \equiv \frac{OZ}{OM}.$$

This ratio is equal to the relation between the costs which would have resulted at the efficient point, *E*, (corresponding to *OZ)* and the total costs incurred at the actual point of production, *M*. Thus, *AE* can be computed according to

$$AE = \frac{C}{C^a} = \frac{g(\boldsymbol{w}, t)}{\sum_{k=1}^{n} w_k \cdot \frac{\partial g(\boldsymbol{w}^\diamond, t)}{\partial(\lambda_i^\diamond w_i)}}. \tag{21}$$

The denominator in the last equality is equal to the denominator of the realized cost shares (20) and, thus, can be directly obtained from the estimation of that system. The numerator is also easy to obtain; due to the linear homogeneity of $g(\cdot)$ in input prices $g(\boldsymbol{w},t)$ can be computed simply by setting all the $\lambda_i^\diamond$'s in the denominator of (21) equal to one. Notice that changes in the relative input prices and in the time index will cause AE to vary over time, yielding estimates of the degree of allocative efficiency for each point of observation.

From (21) it is clear that once the AE measures have been computed we can easily estimate the relative increase in total costs caused by the misallocation of inputs, in spite of the fact that we have no measure of output. The relative increase $(C^a - C)/C$ is simply equal to $(1 - AE)/AE$. Finally, for later reference, we note that the cost-minimizing input demands can be expressed in terms of the $\hat{x}_i$'s, according to

$$x_i = \hat{x}_i \cdot \frac{\dfrac{\partial g(\boldsymbol{w}, t)}{\partial w_i}}{\dfrac{\partial g(\boldsymbol{w}^\diamond, t)}{\partial(\lambda_i^\diamond w_i)}}. \tag{22}$$

4.2. Allocative and technical inefficiency

We now relax the assumption of technical efficiency. In general terms, a producer is defined as technically inefficient if, at a given level of production, he/she can reduce the utilization of any input and still produce the same amount of output. In Fig. 1 above, technical inefficiency is illustrated by the point B, which cannot be technically efficient as it is not on the efficient production surface II'.

Farrell (op. cit.) has proposed a simple measure of the degree of technical efficiency *(TE)*. In terms of the diagram, it is defined as

$$TE \equiv \frac{OM}{OB}.$$

A convenient interpretation of this measure is obtained by considering the difference $1 - TE$ which, by definition, belongs to the open interval $]0,1]$. For the given level of output, $1 - TE$ shows the potential relative decrease in the input utilization when the factor proportions are held constant, i.e. when the relative reduction is constrained to be the same for all inputs. Since TE is defined relative to the factor ray through the origin and the observed point it is a *radial* measure of technical efficiency. Like the Farrell measure of allocative efficiency (AE), TE can also be expressed in terms of total costs. The ratio OM/OB is equal to the total costs associated with the technically efficient (but allocatively inefficient) point M, divided by the total costs incurred at B, the point actually observed. Denoting the total costs actually observed by C^+, the degree of technical inefficiency can be formulated according to

$$TE = \frac{C^a}{C^+}.\text{[23]} \tag{23}$$

An appealing property of the technical efficiency measure TE is that it does not affect the Farrell measure of of allocative efficiency, AE. This is easily seen in Fig. 1. Different degrees of technical efficiency correspond to different locations on the dashed ray through the origin, on or above the isoquant II'. But for all these points the degree of allocative efficiency is the same, namely OZ/OM. This means that degrees of allocative and technical efficiency can be independently computed. The degree of overall efficiency (OE) is simply given by

$$OE = TE \times AE = \frac{C}{C^+}, \tag{24}$$

where the last equality follows from (21) and (23).

Unfortunately, it is not possible to estimate TE directly by means of the system of cost shares. The reason is that the radial specification of technical inefficiency is *input neutral* and hence, like neutral technical change,

[23] Farrell defined TE for a constant returns technology. The extension to more general technologies is due to Førsund and Hjalmarsson (1974, 1979).

has no effect on the input cost shares.[24] This is easily shown formally, as follows. Let $\tilde{x}_i$ denote the demand for input i in the context of both allocative and technical inefficiency. Defining the $\tilde{x}_i$ in terms of the technically efficient (but allocatively inefficient) input demands given by (18), we must have

$$\tilde{x}_i \equiv (1 + \zeta)\cdot\hat{x}_i\,, \qquad \zeta \geq 0, \qquad i = 1,\ldots,n, \tag{25}$$

where ζ represents the common degree of overutilization, implying that

$$TE = (1 + \zeta)^{-1}\,. \tag{26}$$

Further, denote the total costs actually incurred in the context of both allocative and technical inefficiency by C^+. Then, by definition,

$$C^+ \equiv \sum_{i=1}^{n} w_i\tilde{x}_i = (1 + \zeta)\cdot C^a\,. \tag{27}$$

Together with (25), (27) implies that the cost shares $(w_i\tilde{x}_i)/C^+$ are equal to the cost shares prevailing in the context of allocative inefficiency only, i.e. the s_i^a. Notice that this result implies that, in addition to allocative inefficiency, the system (20) also implicitly allows for radial technical inefficiency.[25]

However, to be able to take technical inefficiency *explicitly* into account we have to let it affect the input usage in a *non-radial* fashion, i.e. allow the

[24] This statement is not inconsistent with the demonstration in Section 3.1 that there are conditions when the effects of neutral technical change can be estimated, in spite of the fact that the cost shares are invariant to neutral technical change. What was shown there was that the share equations can generate estimates of neutral technical change *provided* that neutral and non-neutral technical change occur simultaneously. While, in principle, a similar statement could be made about technical inefficiency, a simultaneous analysis of both radial and non-radial technical inefficiency would not be meaningful.

[25] Since the system (7) of input cost shares is a special case of the system (20), this invariance property implies that in the presence of radial technical inefficiency estimation of the system (7) is still valid, and will yield unbiased estimates, although the assumption of cost minimization is violated.

degree of overutilization to vary among the inputs.[26] To this end we will derive a system of input cost shares which takes the combined effects of technical and allocative inefficiency, i.e. overall inefficiency, into account and which includes the system (20) as a special case.

Unfortunately, in the system of cost shares allowing for overall inefficiency it is not possible to separate allocative from technical inefficiency in an unambiguous way. The reason is that the introduction of input-specific degrees of overutilization removes the independence between the measures of technical and allocative efficiency, which is characteristic of the Farrell scheme, cf. Kopp (1981).[27] Provided, however, that we make the assumption that the production technology satisfies *strong free disposability of inputs* (SFDI) the system allowing for overall inefficiency can be combined with the system (20) to yield a Farrell decomposition of the overall inefficiency in accordance with (21), (23) and (24). SFDI implies that when production is taking place at a technically efficient point an increase in the utilization of some input(s) will always result in some, however small, increase in output. As noted by Kopp (op. cit.), most of the functional forms employed in econometric production studies satisfy SFDI. Among them are the CES and the translog; cf. Färe and Lovell (1978) and Kopp and Diewert (1982), respectively.[28] The condition of SFDI ascertains that a given degree of overall

---

[26] Non-radial specifications of technical inefficiency have been considered by Färe (1975) and by Färe and Lovell (1978).

[27] This property does not seem to have been generally recognized in the literature. For instance, in the empirical application of a model allowing for both allocative and non-radial technical inefficiency, Lovell and Sickles (1983) use an estimation method which treats these two types of inefficiency as if they were independent.

[28] Notice that the condition of SFDI is slightly more restrictive than that of free disposability of inputs (FDI), which is fulfilled by all technologies which have a dual representation (cf. footnote 19). An example of a flexible functional form which does not satisfy SFDI globally is the Generalized Leontief. In particular, its special case the (ordinary) Leontief technology fails SFDI everywhere.

efficiency can always be equivalently decomposed into *either* non-radial technical inefficiency and allocative inefficiency *or* radial technical inefficiency and allocative inefficiency (although the measures of allocative inefficiency will differ in the two cases). This property is illustrated in Fig. 2.

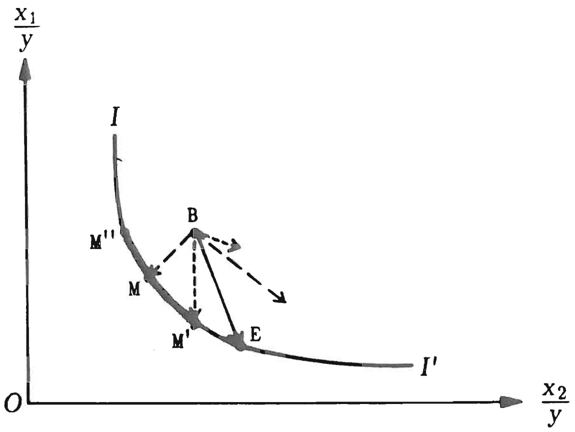

Figure 2. Equivalent decompositions of overall efficiency

The isoquant and the points *B*, *M*, and *E* have been reproduced from Fig. 1. We now make the thought experiment that the producer operating at the point *B* moves to the efficient point, *E*. This movement, illustrated by the solid arrow, can be considered as the sum of two vectors, representing movements towards technical and allocative efficiency, respectively. In principle, the sum can be decomposed in an infinite number of ways. The vector corresponding to the adjustment towards technical efficiency must, however, result in a point on the boldly drawn part of the isoquant whose endpoints coincide with the points M' and M''. This is so because, by definition, technical inefficiency corresponds to *over*utilization of inputs. The movement to a technically efficient point thus cannot involve an increase in the use of any input.

The dashed vectors illustrate the special case in which the adjustment towards technical efficiency is radial, i.e. when (25) holds. The adjustment yielding technical efficiency is represented by the vector from B to M, whereas the other vector is equivalent to the movement from M to E, i.e. the movement for allocative efficiency. Of the dotted vectors the one pointing due south, to M', corresponds to a non-radial adjustment towards technical efficiency where the amount of input 2 is held constant while decreasing the use of input 1. By elimination, the other vector must then show the movement yielding allocative efficiency.

We now proceed to derive a system of input cost shares allowing for both allocative inefficiency and non-radial technical inefficiency. We begin by assuming that the input demands (25) can be equivalently represented according to

$$\tilde{x}_i = f(y)\cdot\frac{\partial g(\boldsymbol{w}^\star, t)}{\partial(\lambda_i^\star w_i)} + f(y)\cdot\mu_i\,, \quad \mu_i \geq 0 \quad i = 1,\ldots,n, \tag{28}$$

where

$$\boldsymbol{w}^\star = (\lambda_1^\star w_1,\ldots,\lambda_{n-1}^\star w_{n-1}, w_n)$$

The last term on the RHS represents the excessive usage of input i. For simplicity, the μ_i's are here taken to be parametrical constants.[29] The excessive input usage is thus assumed to vary between the inputs and to change with the scale of operation. In particular, if returns to scale are constant then the the overutilization is proportional to the level of output.

[29] As parameters, the μ_i's may not be identified for all kinds of functional forms. However, regarding, e.g., the CES and translog functional forms, which we know satisfy SFDI, identification is always possible. Chapter III considers identification in the context of a translog cost function.

Notice that, in general, the first term on the RHS of (28) is not equal to $\hat{x}_i$, given by (18). This is indicated by the use of the superindex $\star$ on the λ_i's instead of the $\diamond$ used in Section 4.1. As discussed above, the reason why the $\lambda_i^\diamond$ and the $\lambda_i^\star$ differ is that the shadow prices associated with radial technical inefficiency are distinct from the shadow prices associated with non-radial technical inefficiency. The partial derivative $\partial g(\boldsymbol{w}^\star,t)/\partial(\lambda_i^\star w_i)$ may thus be either greater or smaller than the partial derivative $\partial g(\boldsymbol{w}^\diamond,t)/\partial(\lambda_i^\diamond w_i)$ which, together with $f(y)$, determines $\hat{x}_i$ according to (18). This means that in addition to being non-negative the μ_i must also fulfill the condition

$$\mu_i \geq \frac{\partial g(\boldsymbol{w}^\diamond, t)}{\partial(\lambda_i^\diamond w_i)} - \frac{\partial g(\boldsymbol{w}^\star, t)}{\partial(\lambda_i^\star w_i)}, \qquad i = 1,\ldots,n, \tag{29}$$

in order to ensure that the inequalities $\tilde{x}_i \geq \hat{x}_i$, $i = 1,\ldots,n,$ hold.

The specification (28) is just one among several possible ways to account for non-neutral technical inefficiency. We have chosen this particular specification because it is simple and because it leads to input cost shares which are independent of y. In the latter respect it differs from related specifications, for example the one used by Lovell and Sickles (1983), which amounts to substituting μ_i for $f(y)\cdot\mu_i$ in (28). Lovell and Sickles' approch yields a model in which techhnical inefficiency declines in relative terms when production is increased. Our formulation does not have this somewhat unappealing property.

Up to a constant of integration, the cost function corresponding to (28) is

$$\tilde{C} \equiv f(y)\cdot g(\boldsymbol{w}^\star,t) + f(y)\cdot \boldsymbol{\mu}'\boldsymbol{w}^\star , \tag{30}$$

where the prime in the last term denotes transposition. It is straightforward to show that for the price vector $\boldsymbol{w}^\star$ the cost function $\tilde{C}$ is regular. Thus,

$\tilde{C}$ can safely be considered to be the dual representation of some production technology and the application of Shephard's lemma to (30), yielding (28), is justified. The total costs actually observed are not given by (30), however. In accordance with the definition given in (27) the total costs actually observed can be expressed as

$$C^{+} = f(y) \cdot \sum_{k=1}^{n} w_k \left[\frac{\partial g(\boldsymbol{w}^{\star}, t)}{\partial(\lambda_k^{\star} w_k)} + \mu_k \right] . \tag{31}$$

If all the μ_k's are equal to zero (and, hence, $\lambda_k^{\star} = \lambda_k^{\diamond}$ for $k = 1,\dots,n$) then (31) reduces to (19), the total cost realized in the context of allocative inefficiency only. Given (28) and (31) the observed input costs shares can be written

$$s_i^{+} \equiv \frac{w_i \cdot \tilde{x}_i}{C^{+}} = \frac{w_i \cdot \dfrac{\partial g(\boldsymbol{w}^{\star}, t)}{\partial(\lambda_i^{\star} w_i)} + w_i \mu_i}{\sum_{k=1}^{n} w_k \cdot \left[\dfrac{\partial g(\boldsymbol{w}^{\star}, t)}{\partial(\lambda_i^{\star} w_i)} + \mu_k \right]}, \qquad i = 1,\dots,n. \tag{32}$$

In the estimation of (32), the $\lambda_i^{\star}$'s should be subjected to the same constraints as the those imposed on the $\lambda_i^{\diamond}$'s in the estimation of the system (20).[30] Concerning the μ_i's the non-negativity restrictions in (28) pose no problem; they can be implemented by means of the same method as the one employed to ensure positive values on the $\lambda_i^{\diamond}$'s and the $\lambda_i^{\star}$'s. The inequality constraints (29) are more difficult to impose, however. The simplest way to proceed is probably to ignore them in a first round estimation. Should a comparison with the estimates obtained from the estimation of (20) reveal that any of the μ_i's violate (29) for some observations, then the lower bound for these parameters can be raised above zero according to $\mu_i \geq \mu_i^1 + \kappa_i$

[30] For clarity, it should be pointed out that "estimation of (20)" is equivalent to "estimation of (32) subject to the constraint $\mu_i = 0$ for $i = 1,\dots,n$". We use the former expression for the obvious reason that it is shorter and simpler.

where μ_i^1 is the first round estimate and κ_i a suitably chosen positive number. It is not certain that the second-round estimates will satisfy (29) either but the inequality can always be made to hold by repeating the procedure.[31]

When both the systems (20) and (32) have been estimated a likelihood ratio test can be performed of the null hypothesis $H_0 : \mu_i = 0,\ i = 1,\ldots,n$. The test of H_0 corresponds to a weak test of technical efficiency – the test is weak in the sense that the system under the null, i.e. (20), is consistent both with technical efficiency and radial technical inefficiency. Rejection of H_0 implies, however, that the production process cannot be technically efficient.

At first, rejection of the hypothesis that all the μ_i's are equal to zero might seem as an implausible outcome. Given that the two decompositions of overall inefficiency, involving radial and non-radial technical inefficiency, respectively, are indeed equivalent, then why should the latter decomposition be preferred to the former? However, this objection fails to recognize that the fact that there exists alternative decompositions of overall inefficiency which are *mathematically* equivalent does not imply that these alternatives are also statistically equivalent, in the sense of providing equally good fit to data. Rather, it is reasonable to expect the more richly parameterized alternative to be preferred to the more parsimonious one. Hence, if the production process is technically inefficient, rejection of H_0 should be a more likely outcome than acceptance.

The estimated versions of the systems (20) and (32), yield an estimate of

[31] That the procedure must eventually produce estimates which satisfy (29) is clear from (28); an increase in the second term on the RHS due to an increase in μ_i must be balanced by a decrease in the first term, brought about by a decrease in the partial derivative. Thus, when the LHS of (29) goes up, the RHS goes down.

TE according to

$$TE = \frac{\sum_{k=1}^{n} w_k \cdot \dfrac{\partial g(\boldsymbol{w}^{\diamond}, t)}{\partial(\lambda_k^{\diamond} w_k)}}{\sum_{k=1}^{n} w_k \cdot \left[\dfrac{\partial g(\boldsymbol{w}^{\star}, t)}{\partial(\lambda_k^{\star} w_k)} + \mu_k\right]}, \tag{33}$$

cf. (23). Like the Farrell measure of the degree of allocative efficiency, (21), TE varies over time in response to changes in the relative input prices and the time index. By (26), ζ, the common degree of overutilization corresponding to neutral technical inefficiency, is given by

$$\zeta = \zeta(\boldsymbol{w}^{\diamond}, \boldsymbol{w}^{\star}, t) = (TE)^{-1} - 1 \,. \tag{34}$$

Thus, ζ is not a constant but a function, determined by the input prices and the time index. Given the estimate of ζ, the input utilization in the context of allocative inefficiency only, i.e. the $\hat{x}_i$'s, can be computed by dividing the actually observed input usage $\tilde{x}_i$, $i = 1,\dots,n$, by $(1 + \zeta)$; cf. (25). Finally, by inserting the so obtained estimates into (22) we obtain estimates of the cost-minimizing input demands, too. Hence, it is possible to compare the input usage actually observed with the cost-minimizing levels of utilization and to compute the minimum total costs, i.e. $C \equiv \Sigma_{k=1}^{n} w_k x_k$, in spite of the presumed lack of output measure.

4.3 Computation of elasticities and effects of technical change

To enable comparisons between the price and substitution elasticities prevailing under cost-minimization, i.e. (4) and (5), and those corresponding to the input utilization actually observed, we show here how the latter elasticities should be computed. Likewise, we consider the effects of technical change corresponding to the $\tilde{x}_i$, $i = 1,\dots,n$.

We first derive the elasticities of substitution. These should be defined in terms of the cost function from which the $\tilde{x}_i$ have been derived. Since $\tilde{x}_i = \partial\tilde{C}/\partial(\lambda_i^\star w_i)$, $i = 1,\dots,n$, this means that the cost function (30) should be used. In order to obtain a formula for the actual elasticities which is analogous to (4) we need the input cost shares which are minimum for this cost function. Denoting these by $\tilde{s}_i$ we obtain

$$\tilde{s}_i \equiv \frac{(\lambda_i^\star w_i)\cdot\dfrac{\partial\tilde{C}}{\partial(\lambda_i^\star w_i)}}{\tilde{C}} = \frac{(\lambda_i^\star w_i)\cdot\left[\dfrac{\partial g(\boldsymbol{w}^\star, t)}{\partial(\lambda_i^\star w_i)} + \mu_i\right]}{g(\boldsymbol{w}^\star, t) + \boldsymbol{\mu}'\boldsymbol{w}^\star}, \tag{35}$$

$i = 1,\dots,n$. To compute the numerator of $\tilde{s}_i$ we simply multiply the estimated numerator of s_i^+ by $\lambda_i^\star$; cf. (32). And, as usual, the denominator is equal to the sum of the numerators. In analogy with (4), the elasticities can be expressed in terms of the $\tilde{s}_i$, according to

$$\tilde{\sigma}_{ij} = \left[\tilde{s}_i\tilde{s}_j + (\lambda_j^\star w_j)\cdot\frac{\partial\tilde{s}_i}{\partial w_j}\right]\cdot(\tilde{s}_i\tilde{s}_j)^{-1}. \tag{36}$$

Because of the proportionality between the shadow prices and the input prices actually observed the price elasticities can be obtained as follows

$$\tilde{\eta}_{ij} \equiv \frac{\partial\tilde{x}_i}{\partial w_j}\frac{w_j}{\tilde{x}_i} = \frac{\partial\tilde{x}_i}{\partial(\lambda_j^\star w_j)}\frac{(\lambda_j^\star w_j)}{\tilde{x}_i} = \tilde{s}_j\tilde{\sigma}_{ij}, \tag{37}$$

where the first equality is due to the chain rule and the second equality follows by analogy with (5); this analogy is justified because the $(\lambda_j^\star w_j)$'s are the prices for which the cost function (30) is defined.

Concerning technical change, its effects on input utilization are given by

$$\tilde{\tau}_{x_i} \equiv \frac{\partial\tilde{x}_i(y,\boldsymbol{w}^\star,t)}{\partial t}\frac{1}{\tilde{x}_i(y,\boldsymbol{w}^\star,t)} = \frac{\partial^2 g(\boldsymbol{w}^\star,t)}{\partial(\lambda_i^\star w_i)\partial t}\left[\frac{\partial g(\boldsymbol{w}^\star,t)}{\partial(\lambda_i^\star w_i)}\right]^{-1}, \quad i = 1,\dots,n \tag{38}$$

Regarding the effects on total costs, two aspects are relevant. On the one hand, it is of interest to consider the influence that technical change has had on the total costs actually observed, i.e.

$$\tau_C^+ \equiv \frac{\partial C^+}{\partial t}\frac{1}{C^+} = \sum_{i=1}^{n} s_i^+ \tilde{\tau}_{x_i} . \tag{39}$$

On the other hand, from the producer's point of view it is relevant to compute the effects of technical change on the total cost function (30), as (30) is the dual representation of the production technology (under the false perception that the input price vector is given by $\boldsymbol{w}^\star$ rather than $\boldsymbol{w}$). This means that in the aggregation of the $\tilde{\tau}_{x_i}$ the $\tilde{s}_i$'s should be used as weights according to

$$\tilde{\tau}_C \equiv \frac{\partial \tilde{C}}{\partial t}\frac{1}{\tilde{C}} = \sum_{i=1}^{n} \tilde{s}_i \tilde{\tau}_{x_i} . \tag{40}$$

Like τ_C [defined by (11)] $\tilde{\tau}_C$ has a dual interpretation: $-\tilde{\tau}_C$ is the rate of technical change in the production function to which (30) is dual. No dual interpretation is possible with respect to τ_C^+, however, since C^+ is not a regular cost function. Finally, by analogy with note 16, the relative changes in s_i^+ and $\tilde{s}_i$ induced by technical change are $\tilde{\tau}_{x_i} - \tau_C^+$ and $\tilde{\tau}_{x_i} - \tilde{\tau}_C$.

5. Summary and concluding comments

What can you learn about a production process for which no output measures are available? This is the question we have tried to answer with the help of duality theory. Our results show that the possibilities to characterize production by means of input data only are indeed much greater than could be expected intuitively.

The fundamental property upon which we base our analysis, i.e. the fact that for a homothetic technology the input cost shares can be completely specified without any information about output, can be found already in Shephard (1953). The importance of this result for applied production theory seems not to have been recognized, however, which is surprising considering the tremendous growth that has since then occurred in the production of services, where the output measurement problems are especially severe. It is significant that Hulten's (1984) study of productivity change in the public sector, the only previous attempt to characterize a production process econometrically without explicit measures of the price or quantity of output, did not escape the output measurement problem by considering the input cost shares. Instead, Hulten chose to regard communities as generalized households, thereby making it possible to apply the analytical apparatus of the household production model to the production of public services.

In contrast to Hulten's framework, our method can be applied to any production activity. Moreover, our analysis goes beyond Hulten's in that it is not limited to the issue of estimating productivity growth. We show that given a homothetic technology, knowledge of input prices and input cost shares makes it possible to estimate elasticities of substitution and factor demand, analyze productivity effects of technical change, and study (deviations from) efficiency in production.

Concerning the relationship between technical change and productivity growth, we show that the relative effects of technical change on total costs always can be estimated but that these correspond to estimates of the dual rate of growth in total factor productivity (TFP) only if constant returns to scale are assumed, as in Hulten's study. If, instead, homogeneity of degree $r \neq 1$ is assumed the rate of growth in TFP can be estimated up to an initial condition or bench-mark value, while homotheticity allows only the sign of the TFP growth rates to be determined.

We also demonstrate how possible deviations from cost minimization can be taken into account parametrically. Here, we make use of the fact that for a large class of technologies overall inefficiency can be decomposed either into independent measures of *radial* technical inefficiency and allocative inefficiency according to Farrell (1957), or into two interdependent measures of *non-radial* technical inefficiency and allocative inefficiency.[32] Since the input cost shares are invariant with respect to radial technical inefficiency, the Farrell decomposition results in a share system which can take allocative inefficiency explicitly into account but which only allows for (radial) technical inefficiency implicitly, making it impossible to quantify the latter. The second decomposition, on the other hand, yields a system of cost shares by means of which overall inefficiency can be measured but which cannot separate clearly between technical and allocative inefficiency. We show that by estimating *two* share systems, one for each decomposition, Farrell measures of technical, allocative, and overall inefficiency can be obtained. Moreover, the increases in total costs brought about by the inefficiencies can be estimated as well as the cost-minimizing input demands, in spite of the presumed lack of output data.

Finally, it should be mentioned that although we have here performed the analysis in terms of static equilibrium cost models it can be extended to allow for the possibility that some of the factors of production, notably capital, may be fixed in the short run. For instance, it should be possible to use a dynamic cost of adjustment model of the type employed by Berndt and Hesse (1986) as a starting point. Since a typical feature of many service industries is that the bulk of their capital input is in the form of structures this extension should be an important one.

[32] The measures of allocative inefficiency will, of course, differ between the two alternative decompositions.

References

Allen, R.G.D., 1959, *Mathematical Analysis for Economists*, London: Macmillan Press.

Atkinson, S.E. and R. Halvorsen, 1984, Parametric Efficiency Tests, Economies of Scale and Input Demand in U.S. Electric Power Generation, *International Economic Review* 25, pp. 647 – 662.

Atkinson, S.E. and R. Halvorsen, 1986, The Relative Efficiency of Public and Private Firms in a Regulated Environmement: The Case of U.S. Electric Utilities, *Journal of Public Economics* 29, pp. 281 – 294.

Barten A.P., 1969, Maximum Likelihood Estimation of a Complete System of Demand Equations, *European Economic Review* 1, pp. 7 – 73.

Berndt, E.R. and D.M. Hesse, 1986, Measuring and Assessing Capacity Utilization in the Manufacturing Sectors of Nine OECD Countries, *European Economic Review* 30, pp. 961 – 989.

Berndt, E.R. and M.S. Khaled, 1979, Parametric Productivity Measurement and Choice Among Flexible Functional Forms, *Journal of Political Economy* 87, pp. 1220 – 1245.

Binswanger, H.P., 1974, The Measurement of Technical Change Biases with Many Factors of Production, *American Economic Review* 64, pp. 964 –76.

Blackorby, C., C.A.K. Lovell, and M.C. Thursby, 1976, Extended Hicks Neutral Technical Change, *Economic Journal* 86, pp. 845 – 852.

Byrnes, P., S. Grosskopf, and K. Hayes, 1986, Efficiency and Ownership: Further Evidence, *Review of Economics and Statistics* 68, pp. 337 – 341.

Diewert, W.E., 1971, An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function, *Journal of Political Economy* 79, pp. 481 – 507.

Eakin, B.K. and T.J. Kniesner, 1988, Estimating a Non-minimum Cost Function for Hospitals, *Southern Economic Journal* 54, pp. 583 – 597.

Färe, R., 1974, On Linear Expansion Paths and Homothetic Production Functions, in W. Eichhorn, R. Henn, O. Opitz, and R.W. Shephard, eds., *Production Theory*, Heidelberg: Springer-Verlag.

Färe, R., 1975, Efficiency and the Production Function, *Zeitschrift für Nationalökonomie* 35, pp. 317 – 324.

Färe, R., S. Grosskopf, and C.A.K. Lovell, 1985, *The Measurement of Efficiency of Production*, Boston: Kluwer Nijhoff.

Färe, R. and C.A.K. Lovell, 1978, Measuring the Technical Efficiency of Production, *Journal of Economic Theory* 19, pp. 150 – 162.

Farrell, M.J., 1957, The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society, Series A* 120, pp. 253 – 281.

Førsund, F.R., 1975, The Homothetic Production Function, *Swedish Journal of Economics* 77, pp. 234 – 244.

Førsund, F.R. and L. Hjalmarsson, 1974, On the Measurement of Productive Efficiency, *Swedish Journal of Economics* 76, pp. 141 – 154.

Førsund, F.R. and L. Hjalmarsson, 1979, Generalized Farrell Measures of Efficiency: An Application to Milk Processing in Swedish Dairy Plants, *Economic Journal* 89, pp. 294 – 315.

Fuss, M. and D. McFadden, 1978, Flexibility versus Efficiency in *Ex Ante* Plant Design, in M. Fuss and D. McFadden eds., *Production Economics: A Dual Approach to Theory and Applications,* Vol. I, Amsterdam: North-Holland.

Hulten, C.R., 1984, Productivity Change in State and Local Governments, *Review of Economics and Statistics* 66, pp. 256 – 266.

Jorgenson, D.J. and Z. Griliches, 1967, The Explanation of Productivity Change, *Review of Economic Studies* 34, pp. 249 – 282.

Kopp, R.J., 1981, Measuring the Technical Efficiency of Production: A Comment, *Journal of Economic Theory* 25, pp. 450 – 452.

Kopp, R.J. and W.E. Diewert, 1982, The Decomposition of Frontier Cost Function Deviations into Measures of Technical and Allocative Efficiency, *Journal of Econometrics* 19, pp. 319 – 331.

Lau, L.J., 1978, Testing and Imposing Monotonicity, Convexity, and Quasi-convexity Constraints, in M. Fuss and D. McFadden, eds., *Production Economics: A Dual Approach to Theory and Applications*, Vol. I, Amsterdam: North-Holland.

Lau, L.J. and P.A. Yotopoulos, 1971, A Test for Relative Efficiency and Application to Indian Agriculture, *American Economic Review* 61, 94-109.

Lovell, C.A.K. and R.C. Sickles, 1983, Testing Efficiency Hypotheses in Joint Production: A Parametric Approach, *Review of Economics and Statistics* 65, pp. 51 – 58.

Nadiri, I. and M. Schankerman, 1980, The Structure of Production, Technological Change, and the Rate of Growth of Total Factor Productivity in the U.S. Bell system, in T.G. Cowing and R.E. Stevenson, eds., *Productivity Measurement in Regulated Industries*, New York: Academic Press.

Ohta, M., 1974, A Note on the Duality between Production and Cost Functions: Rate of Returns to Scale and Rate of Technical Progress, *Economic Studies Quarterly* 25, pp. 63 – 65.

Parsons, D.J., C.G. Gotlieb, and M. Denny, 1992, Productivity and Computers in Canadian Banking, forthcoming in *Journal of Productivity Analysis* 3.

Rhoades, S.A. and A.P. White, 1984, Output in Relation to Labor Input in the Banking and Savings and Loan Industries 1927 – 1979, *Journal of Banking and Finance* 8, pp. 119 – 130.

Shephard, R.W., 1953, *Cost and Production Functions*, Princeton, New Jersey: Princeton University Press.

Toda, Y., 1976, Estimation of a Cost Function when the Cost is Not Minimum: The Case of Soviet Manufacturing Industries 1958 – 1971, *Review of Economics and Statistics* 58, pp. 259 – 268.

Toda, Y., 1977, Substitutability and Price Distortion in the Demand for Factors of Production: An Empirical Estimation, *Applied Economics* 9, pp. 203 – 217.

Uzawa, H., 1962, Production Functions with Constant Elasticities of Substitution, *Review of Economic Studies* 29, pp. 291 – 299.

CHAPTER III

Identification of technical and allocative inefficiencies in the absence of output data

1. Introduction

It is well known that for homothetic production technologies – i.e. technologies which have the property that the optimal factor proportions do not vary with the level of production – the input cost shares are independent of the output level. Accordingly, if one assumes homotheticity and models the production technology by means of the cost function one can, e.g., estimate elasticities of substitution among inputs and the effects of (non-neutral) technical change without any information about output. This, of course, is a very useful property in situations where it is difficult to construct reliable output measures. Perhaps the most important example concerns the production of public services.[1]

However, concerning public production in particular, the validity of taking a dual approach can be questioned. The underlying assumption of cost minimization does not agree with common opinions about misallocation of resources and excessive input usage in the public sector. This raises the issue of whether it is possible to model inefficiencies in production within a dual framework and, if so, if it is possible to do it in such a way that the inefficiencies will be reflected in the cost shares.

---

[1] The homotheticity assumption is often considered to be quite restrictive. In Chapter II some arguments are provided, however, according to which homotheticity should be more easily justified in the context of service production than in the production of goods.

Regarding allocative inefficiency, one could say that this question has already been answered in the affirmative. Confining the attention to models in which inefficiencies are taken into account parametrically, studies by, among others, Toda (1976, 1977) and by Atkinson and Halvorsen (1984, 1986) can be mentioned. In these analyses allocative inefficiency is allowed for by modeling the producer as minimizing costs subject to another set of (relative) input prices than those actually observed. Specifically, it is assumed that the producer behaves as if the relative input prices subject to which he/she optimizes are proportional to the observed relative input prices. By specifying the cost function in terms of the hypothetical, rather than the actual, input prices one can then apply the usual results from duality theory.

In all the published studies using this technique, output measures have been available and so there has been no need to estimate merely the cost share equations. In Chapter II of this book it has been argued, though, that, in addition to yielding estimates of the allocative inefficiency parameters, estimation of the system of input cost shares would also make it possible to compute estimates of the increases in total costs caused by the allocative inefficiency, as well as its effects on input demands. The intuition behind this claim is that since allocative inefficiency relates to the input proportions, which are independent of output for homothetic technologies, the evaluation of the effects of allocative inefficiency should not require output information.

However, returning to common opinion, technical inefficiency, i.e. excessive input usage, is probably a more serious problem than allocative inefficiency. The problem of modeling technical inefficiency can be solved by means of an approach similar in spirit to the one used for modeling allocative inefficiency. To see this, note that the technology available to the producer determines the minimum input requirements associated with every output level. Next, consider a second technology which is dominated by the first in the sense that is characterized by minimum input requirements which, for every given level of output, are at least equal to those of the first technology. Technical inefficiency can be allowed for by modeling the producer as trying to minimize costs subject to the inferior, rather than the superior, technology. Note that the inferior technology is contained in the superior technology and that in the case of technical efficiency the two will be equal. That neither of the two technologies can be observed does not constitute a problem; it is sufficient that the cost functions representing them can be estimated.

Of course, what this says is merely that the cost function can be generalized to allow for technical as well as allocative inefficiency. If, in addition, the technical inefficiency is to have any effect on the input cost shares there is an additional requirement, namely that technical inefficiency be *non-radial.* This means that, in percentage terms, the excessive usage varies over inputs.[2] While this may seem as a very natural assumption to make, the seminal work of Farrell (1957) has created a strong tradition of modeling technical inefficiency as radial (input neutral). If the Farrell route is taken in conjunction with a dual approach it is necessary to complement the cost share equations by the cost function in order to obtain estimates of technical inefficiency.[3]

Thus, the lack of output data presumed here implies that technical inefficiency has to be modeled as non-radial. The important question, which provides the basic motivation for this chapter, can then be put as follows. Is it possible to model non-radial technical inefficiency parametrically such that it (i) affects the input cost shares without making them dependent on the level of output, and (ii) allows the effects of this inefficiency on input

---

[2] For further discussion, see Färe and Lovell (1978).

[3] Examples of studies where allocative inefficiency and radial technical inefficiency are modeled by means of the input cost shares and the cost function are Kopp and Diewert (1982) and Ferrier and Lovell (1990).

demands and total costs to be inferred from the estimation of the system of cost shares alone?

In Chapter II a specification of non-radial technical inefficiency was suggested which, it was argued, would make it possible to answer (i) in the affirmative, at least for some types of cost functions. Without a detailed discussion of whether the relevant parameters could be statistically identified it was conjectured that condition (ii) could be satisfied, too. As technical inefficiency seems to be directly related to the amount of output produced these claims certainly are counterintuitive. Moreover, statistical considerations also give rise to doubts. While the system of input cost shares contains $n - 1$ independent equations where $n = \#\text{inputs}$ (as the shares sum identically to 1) the parametric modeling of non-neutral technical inefficiency increases the number of parameters by n. [4] Can these really all be identified?

The idea of this paper is to show that (at least) one cost function — the translog — can be generalized to allow for allocative inefficiency and non-neutral technical inefficiency in such a way that estimation of the corresponding system of cost share equations will yield estimates of both the extra costs induced by these inefficiencies and of their effects on input demands. For analytical tractability, attention is confined to the case when the number of inputs is (less than or) equal to 3.

In Section 2 the system of input cost shares is derived. Section 3 shows how to evaluate the effects of allocative and technical efficiency on input demands, total costs and cost shares, given that the model's parameters are identified. In Section 4 the econometric model is specified. Sections 5 and 6 adress the identification problem. A Full Information Maximum Likelihood

[4] This is in contrast to the modeling of allocative inefficiency. Since allocative inefficiency relates to the relative input prices, it can be modeled by means of just $n-1$ parameters.

(FIML) estimator is considered and it is demonstrated that under very weak conditions the information matrix has full rank, implying that all parameters are statistically identified. Concluding comments are given in Section 7.

2. The system of input cost shares

The starting point of the analysis is a homothetic cost function

$$C(y,\boldsymbol{w},t) = f(y,t)\cdot g(\boldsymbol{w},t) \ , \tag{1}$$

where y denotes the level of output, $\boldsymbol{w}$ is the $n \times 1$ vector of input prices and t a time index representing the state of technology. The function $f(y,t)$ which is strictly positive and strictly increasing in y determines the scaling properties of the technology and the effects of Hicks-neutral technical change. In the present context there is no need to consider an explicit form for $f(y,t)$.

The cost function is assumed to be of the translog form.[5] For $n = 3$

$$g(\boldsymbol{w},t) = \exp\left[\sum_{i=1}^{3}\alpha_i\cdot\ln w_i + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\gamma_{ij}\cdot\ln w_i \ln w_j + \frac{1}{2}\sum_{i=1}^{3}\tau_i\cdot t\ln w_i\right] . \tag{2}$$

where the parameters to be estimated are denoted by Greek letters. It might be noted that non-neutral technical change is allowed for, through the τ_i's .

Linear homogeneity and symmetry impose the following constraints on the parameters

$$\begin{aligned} &\sum_{i=1}^{3}\alpha_i = 1, \qquad \gamma_{ij} = \gamma_{ji} \ , \ j \neq i \\ &\sum_{j=1}^{3}\gamma_{ij} = \sum_{i=1}^{3}\gamma_{ij} = \sum_{i=1}^{3}\tau_i = 0 \ . \end{aligned} \tag{3}$$

[5] The translog has been chosen because of its flexibility, cf. Guilkey, Lovell and Sickles (1983), and because it allows the overall inefficiency (technical plus allocative) to be decomposed in two equivalent ways; cf. Section 3.

To allow for deviations from cost minimization we consider the following generalization of (1)

$$\tilde{C}(y,\boldsymbol{w}^{\star},t) = f(y,t)\cdot[g(\boldsymbol{w}^{\star},t) + h(\boldsymbol{w}^{\star})] \tag{4}$$

where

$$\boldsymbol{w}^{\star} = (w_1^{\star}, w_2^{\star}, w_3^{\star})' = (\lambda_1 w_1, \lambda_2 w_2, \lambda_3 w_3)' , \tag{5}$$

and

$$h(\boldsymbol{w}^{\star}) = \sum_{i=1}^{3} \mu_i \cdot w_i^{\star} . \tag{6}$$

Deviations from cost minimization due to allocative inefficiency are taken into account by the substitution of shadow input prices $(\lambda_i w_i)$, $i = 1,2,3$, for the input prices actually observed, i.e. the w_i's. For simplicity, the λ_i's are taken to be parametric constants.[6] Accordingly, they can only capture the average, systematic, part of allocative inefficiency. If there is no allocative inefficiency then $\lambda_i = 1$ for $i = 1,2,3$. As only relative prices matter and the cost function is well defined only for strictly positive prices the following constraints can be imposed on the λ_i's without loss of generality

$$\lambda_3 = 1 , \tag{7a}$$

$$\lambda_i > 0 , \qquad i = 1, 2 . \tag{7b}$$

In addition to these natural constraints, the following restriction will be imposed:

$$\lambda_1 = \lambda_2 = \lambda . \tag{7c}$$

This restriction greatly simplifies the calculations in Section 6 without

---

[6] Compared to Chapter II there is a notational difference here. In Chapter II the λ_i's were superindexed by either $\diamond$ or $\star$ in order to distinguish between the values taken on by these parameters under different model specifications. Since this distinction is unimportant in the present context the superindices have been dropped.

impinging on the essential feature of the model, namely, that allocative inefficiency and technical inefficiency are considered simultaneously. As the parametric approach to modeling allocative inefficiency has been extensively discussed in the literature there is no need to dwell any further upon it here.[7]

Technical inefficiency is taken into account by means of the function $h(\boldsymbol{w}^\star)$. For reasons to be explained shortly the parameters μ_i, $i = 1, 2, 3$, in this function are constrained to be non-negative, i.e.

$$\mu_i \geq 0\,, \qquad i = 1, 2, 3\,. \tag{8}$$

Before considering the specification of technical inefficiency more closely, note that if the cost function (1) is regular then so is (4).[8] This means that the demand for input i in the presence of allocative and technical inefficiency can be derived by means of Shephard's lemma, i.e. by partial differentiation of (4) with respect to w_i. Denoting this demand by $\tilde{x}_i$ we have:

$$\begin{aligned} \tilde{x}_i \equiv \partial\tilde{C}\,/\partial w_i^\star = f(y,t)\cdot[g(\boldsymbol{w}^\star,t)\cdot(\alpha_i + \sum_{j=1}^{3}\gamma_{ij}\cdot\ln w_j^\star + \tau_i\cdot\tfrac{t}{2})/w_i^\star\,] \\ + f(y,t)\cdot\mu_i\,, \qquad i = 1, 2, 3. \end{aligned} \tag{9}$$

The term $f(y,t)\cdot\mu_i$ measures the (average) amount of technical inefficiency,

---

[7] See, e.g., Lau and Yotopoulos (1971), Toda (1976, 1977), and Atkinson and Halvorsen (1984, 1986).

[8] A cost function is regular if it is non-decreasing in output and in the input prices, and linearly homogeneous and concave in the price vector; see, e.g., Diewert (1971). The postulated properties of the function f and the constraints (3) are not sufficient to ascertain that the cost function (1) is regular — they do not guarantee that costs are non-decrasing and concave in $\boldsymbol{w}$. However, if (1) is indeed regular then (4) must be regular, too. This can easily be seen by first considering the situation when $\mu_i = 0 \;\forall\, i$. In this case it follows immediately that (4) is regular. Allowing at least one of the μ_i's to be strictly positive has the effect of adding a term to the total cost function which does not violate any of the regularity conditions. As the regularity properties are preserved under addition, the regularity of (4) in the general case follows.

i.e. excessive usage, with respect to input i. The fact that, by definition, technical inefficiency cannot be negative motivates the constraints (8).

It can be seen that technical inefficiency is assumed to be input specific or *non-radial*. Moreover, the factor $f(y,t)$ allows the technical inefficiencies to vary over time and with the level of output. This specification, suggested in Chapter II, differs from the formulation of non-radial technical inefficiency employed by Lovell and Sickles (1983), according to which the inefficiencies are assumed to be constant over time and independent of the level of output.[9] To provide some intuition for (9), note that if $f(y,t) = y$, i.e. if there are constant returns to scale, then the effect of technical inefficiency is to increase the input/output – coefficient of input i by the constant μ_i.

If there are neither allocative nor technical inefficiencies then the $\tilde{x}_i$'s specialize into the input demands under cost minimization, $x_i \equiv \partial C/\partial w_i$, $i = 1, 2, 3$. While, in general, $\tilde{x}_i$ can be either greater or smaller than x_i for a particular i the total costs incurred given input demands $\tilde{x}_i$ and input prices w_i, $i = 1, 2, 3$, will always be at least as large as $C(y,\boldsymbol{w},t)$, provided that $C(y,\boldsymbol{w},t)$ is regular.[10] Denoting the total costs in the context of both

---

[9] Further, in contrast to the specification used here, the one used by Lovell and Sickles (op. cit.) makes the cost shares dependent on the level of output, even if the technology is homothetic.

[10] This can be shown as follows. Expand $C(y,\boldsymbol{w},t)$ around $\boldsymbol{w}^\star$ according to
$$C(y,\boldsymbol{w},t) = C(y,\boldsymbol{w}^\star,t) + \mathrm{d}[C(y,\boldsymbol{w}^\star,t)](\boldsymbol{w}-\boldsymbol{w}^\star) + (\boldsymbol{w}-\boldsymbol{w}^\star)'\mathrm{H}[C(y,\boldsymbol{u},t)](\boldsymbol{w}-\boldsymbol{w}^\star)$$
where $\mathrm{d}[\cdot]$ is the vector of first order partial derivatives, evaluated at $\boldsymbol{w}^\star$, and $\mathrm{H}[\cdot]$ the Hessian matrix, evaluated at a point $\boldsymbol{u}$. The linear homogeneity of C in input prices implies that
$$C(y,\boldsymbol{w}^\star,t) + \mathrm{d}[C(y,\boldsymbol{w}^\star,t)](\boldsymbol{w}-\boldsymbol{w}^\star) = \mathrm{d}[C(y,\boldsymbol{w}^\star,t)]\boldsymbol{w} = C^+\big|_{\boldsymbol{\mu}=\mathbf{0}},$$
where the last equality follows from the definition in (10). Thus, to prove the claim for the case when there is no technical inefficiency it is sufficient to show that the last term in the expansion is non-positive. But this follows directly from the concavity of C in input prices. Finally, allowing for technical inefficiency just strengthens the claim as it means the addition of a strictly positive term to the total costs incurred in the case of allocative inefficiency only.

allocative and technical inefficiency by C^+ we have

$$C^+ \equiv \sum_{i=1}^{3} w_i \tilde{x}_i$$
$$= f(y,t)\cdot\left[[g(\boldsymbol{w}^\star,t)\cdot\sum_{i=1}^{3}(\alpha_i + \sum_{j=1}^{3}\gamma_{ij}\cdot\ln w_j^\star + \tau_i\cdot\frac{t}{2})/\lambda_i] + \sum_{i=1}^{3} w_i\mu_i\right], \quad (10)$$

implying the following cost shares

$$s_i^+ \equiv w_i\tilde{x}_i/C^+$$
$$= \frac{g(\boldsymbol{w}^\star,t)\cdot[\sum_{j=1}^{3}\ln\lambda_j(\gamma_{ij}/\lambda_i) + s_i/\lambda_i] + w_i\mu_i}{g(\boldsymbol{w}^\star,t)\cdot\left[\sum_{i=1}^{3}[\sum_{j=1}^{3}\ln\lambda_j(\gamma_{ij}/\lambda_i) + s_i/\lambda_i]\right] + \sum_{i=1}^{3} w_i\mu_i} \quad (11)$$

for $i = 1, 2, 3,$ where s_i is the cost-minimizing cost share, i.e.

$$s_i \equiv \partial\ln C(y,\boldsymbol{w},t)/\partial\ln w_i = \alpha_i + \sum_{j=1}^{3}\gamma_{ij}\cdot\ln w_j + \tau_i\cdot\frac{t}{2}. \quad (12)$$

Using the property that $\Sigma_{i=1}^{3} s_i = 1$ one can see that if there are neither allocative nor technical inefficiencies – implying that $\lambda_i = 1$ and $\mu_i = 0$ for all i – then $s_i^+ = s_i$ for $i = 1, 2, 3$. Since, in contrast to the $\tilde{x}_i$'s and C^+, the s_i^+'s are independent of the level of output, y, the system of input cost shares can be estimated by means of input data only, which are presumed to be available.

3. Evaluating the effects of inefficiency

Technical and allocative inefficiencies affect input demands and thereby total costs and input cost shares. Using (9), the relation between the cost-minimizing demand for input i and the demand in the presence of both technical and allocative inefficiency can be written

$$\frac{\tilde{x}_i}{x_i} = \frac{g(\boldsymbol{w}^\star, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j^\star + \tau_i\cdot\frac{t}{2})/w_i^\star + \mu_i}{g(\boldsymbol{w}, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j + \tau_i\cdot\frac{t}{2})/w_i}, \tag{13}$$

for $i = 1, 2, 3$. Thus, by solving (13) for x_i one obtains an estimate of the costminimizing level of input i, in spite of the fact that this demand depends on the unknown level of output. Further, given the x_i's, the minimum total costs can be calculated by means of the definition $C \equiv \Sigma_{i=1}^{3} w_i x_i$ and, finally, the costminimizing cost shares are given by $s_i = w_i x_i / C$, $i = 1, 2, 3$. Accordingly, the actually observed total costs, input demands, and cost shares, C^+, $\tilde{x}_i$ and s_i^+, $i = 1, 2, 3$, can all be compared to the corresponding variables under cost minimization.[11]

The measure (13) of overall inefficiency can be multiplicatively decomposed into two components, attributable to technical and allocative inefficiency, respectively. The technical inefficiency component is given by

$$\frac{\tilde{x}_i}{\tilde{x}_i\big|_{\mu_i=0}} = \frac{g(\boldsymbol{w}^\star, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j^\star + \tau_i\cdot\frac{t}{2})/w_i^\star + \mu_i}{g(\boldsymbol{w}^\star, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j^\star + \tau_i\cdot\frac{t}{2})/w_i^\star} \tag{14}$$

which can be solved for the input demand $\tilde{x}_i\big|_{\mu_i=0}$ prevailing when there is allocative inefficiency but no technical inefficiency. Finally, the component corresponding to allocative inefficiency equals

$$\frac{\tilde{x}_i\big|_{\mu_i=0}}{x_i} = \frac{g(\boldsymbol{w}^\star, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j^\star + \tau_i\cdot\frac{t}{2})/w_i^\star}{g(\boldsymbol{w}, t)\cdot(\alpha_i + \Sigma_{j=1}^{3} \gamma_{ij}\cdot \ln w_j + \tau_i\cdot\frac{t}{2})/w_i}. \tag{15}$$

From the estimates $\tilde{x}_i\big|_{\mu_i=0}$, $i = 1, 2, 3$, of the input demands in the context

---

[11] Of course, comparisons can be made over time, too, although the suppression of the variables' time indices has the consequence of not making this possibility explicit.

of allocative inefficiency only, the corresponding estimates of total costs can be computed according to $C^{+}\big|_{\mu=0} = \Sigma_{i=1}^{3}\ w_i \cdot \tilde{x}_i\big|_{\mu_i=0}$ and the associated input cost shares can then be calculated as $s_i^{+}\big|_{\mu=0} = \left[w_i \cdot \tilde{x}_i\big|_{\mu_i=0}\right] / C^{+}\big|_{\mu=0}$.

Note that (13) is the product of (14) and (15). By construction, the corresponding multiplicative decomposition of total costs has the same property, i.e. $C^{+}/C = (C^{+}/C^{+}\big|_{\mu=0}) \times (C^{+}\big|_{\mu=0}/C)$. In terms of an isoquant diagram in input/output-space, the technical inefficiency components $(\tilde{x}_i \;/\; \tilde{x}_i\big|_{\mu_i=0})$ and $(C^{+}/C^{+}\big|_{\mu=0})$ measure the volume and cost effects from production taking place off (above) the isoquant rather than on it. The allocative inefficiency components $(\tilde{x}_i\big|_{\mu_i=0}/x_i)$ and $(C^{+}\big|_{\mu=0}/C)$ show how input demands are affected and how much costs are increased, relative to the minimum costs, when production is taking place on the isoquant but not at the optimum location.

While the interpretation of the measures $\tilde{x}_i/x_i$, $i = 1, 2, 3$, and C^{+}/C of overall inefficiency is straightforward, Kopp (1981) has pointed out that the decompositions of these into measures of technical and allocative inefficiency are not unique. This is due to the non-radial specification of technical inefficiency used here. When technical inefficiency is specified as radial, the technical inefficiency component corresponds to a specific movement from the technically inefficient point in input/output-space to the isoquant, namely along the ray which leaves the input proportions unaffected. The fact that the technical inefficiency measure corresponds to a single point on the isoquant implies that there can be only one measure of allocative inefficiency. By contrast, when technical inefficiency is specified as non-radial there is no restriction imposed on the movement from the technically inefficient point to the isoquant. Depending upon which way is chosen (i.e. depending on the parameters μ_i, $i = 1, 2, 3$) the technical inefficiency measure will differ and so will the measure of allocative inefficiency. That is to say, the price one has to pay for allowing technical inefficiency to be non-radial is that the technical and allocative inefficiency components no longer are independent, as they are under the radial specification.

Kopp (op. cit.) also notes, however, that for technologies which satisfy strong free disposability of inputs (SFDI) a given overall inefficiency can always be decomposed using either a radial or a non-radial specification of technical inefficiency.[12] The translog satisfies SFDI ; in fact that was one of the reasons for choosing it here. Accordingly, the decompositions above can always be transformed to decompositions based on a radial specification of technical inefficiency.

4. The econometric model

The econometric model comprises the first two cost shares. The reason why the 3rd cost share is not included is that the complete system of input cost shares is singular. Since maximum likelihood is the estimation method that will be considered here the estimates will be invariant with respect to the choice of equation to be left out in the estimation; cf. Barten (1969). However, the exclusion of the 3rd share means that identification of the model cannot be ascertained unless the constraints (3) are imposed in the estimation.[13] Together with the constraints (7) and (8), this implies that the vector of unrestricted parameters, which will be denoted $\boldsymbol{\theta}$, can be

---

[12] Strong free disposability of inputs implies that when production is taking place at a technically efficient point an increase in the utilization of some input(s) will always result in some, however small, increase in output.

[13] The constraints (3) can be relaxed, and thus tested, only if the (complete) cost function is estimated. Due to the presumed lack of output data that is not possible in the present context, however.

partitioned according to

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_3 \\ \theta_4 \\ \boldsymbol{\theta}_5 \end{bmatrix} = \begin{bmatrix} (\alpha_1, \alpha_2)' \\ (\gamma_{11}, \gamma_{12}, \gamma_{22})' \\ (\tau_1, \tau_2)' \\ \beta \\ (\nu_1, \nu_2, \nu_3)' \end{bmatrix} . \tag{16}$$

The inefficiency parameters λ and μ_i, $i = 1, 2, 3,$ are exponential functions of β and the ν_i, respectively, i.e.:

$$\lambda = \exp(\beta) , \tag{17}$$

$$\mu_i = \exp(\nu_i) , \quad i = 1, 2, 3, \tag{18}$$

so as to ascertain that λ and the μ_i are non-negative, in accordance with (7) and (8).[14] There are, of course, a number of alternatives to the exponential transformation; the quadratic is perhaps the most obvious one. However, like, e.g., hyperbolic transformations the quadratic one has the disadvantage that it implies multiple optimas to the likelihood function. For this reason, monotonic transformations, like the exponential, are preferable.

Later on, it will be useful to aggregate the partition of $\boldsymbol{\theta}$ according to

$$\boldsymbol{\theta} = (\boldsymbol{\kappa}' , \boldsymbol{\nu}')' ; \quad \boldsymbol{\kappa} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3', \theta_4)' , \quad \boldsymbol{\nu} = \boldsymbol{\theta}_5. \tag{16a}$$

To formulate the econometric model, denote the vector of observed cost shares for the 2 first input cost shares in period t by $\boldsymbol{r}_{t\cdot} \equiv (r_{t1}, r_{t2})$. The corresponding vector of predicted cost shares is given by $\boldsymbol{s}^+_{t\cdot} \equiv (s^+_{t1}, s^+_{t2})$, where the s^+_{ti}'s are defined in accordance with (11) and (12); a detailed specification of $\boldsymbol{s}^+_{t\cdot}$ in matrix terms is given below. Finally, denote by $\boldsymbol{u}_{t\cdot}$ a 2×1 vector of disturbances. The model can then be written in (row)

[14] Actually, λ is to be strictly positive; cf. (7b). However, the probability of λ being exactly equal to zero should be extremely low as it implies that the cost shares are independent of the prices of inputs 1 and 2. Thus, for simplicity, only non-negativity has been imposed. Strict positivity is easily ascertained, however, by adding a small positive constant to the RHS of (17).

vector form according to

$$\boldsymbol{r}_{t\cdot} = \boldsymbol{s}^{+}_{t\cdot}(\boldsymbol{\theta}) + \boldsymbol{u}_{t\cdot}\,, \qquad t = 1,\dots,T \tag{19a}$$

where

$$\boldsymbol{u}'_{t\cdot} \sim N(\mathbf{0}_2\,, \boldsymbol{\Omega})\,, \quad E(\boldsymbol{u}'_{t\cdot}\boldsymbol{u}_{s\cdot}) = \mathbf{0}_{2,2} \quad \text{for} \quad s \neq t \tag{19b}$$

and $\mathbf{0}_2$ and $\mathbf{0}_{2,2}$ denote the 2×1 vector of zeros and the 2×2 zero matrix. The 2×2 covariance matrix $\boldsymbol{\Omega}$ is taken to be positive definite.

There are, at least, two ways to motivate the presence of the disturbance terms. One reason is that the producer may be committing errors when minimizing total shadow costs.[15] Secondly, allocative and technical inefficiency are here modeled in a very simplistic and stylized fashion. The parametric specifications may capture the systematic parts of the inefficiencies but there is probably some non-systematic variation, too. Such variations could also have the effect of making the observed cost shares different from the s^{+}_{i} 's.

Concerning the distributional assumption (19b), it is primarily motivated by simplicity. In general, nothing can be said about the disturbances except that they can be both positive and negative. A symmetric distribution therefore seems natural. Regarding the variances and covariances of the disturbances, they may well be heteroscedastic and/or autocorrelated in addition to being contemporaneously correlated. However, as the assumption of a time-independent $\boldsymbol{\Omega}$ will yield consistent estimates even if the residuals are heteroscedastic and/or autocorrelated it should do as a first approximation.[16]

---

[15] Notice that from the producer's perspective the minimum total shadow costs are given by (4).

[16] One potential drawback with the normality assumption is that it may yield predicted cost shares which are either negative or larger than one. From this point of view the Dirichlet distribution, which automatically limits the shares to the unit simplex, is preferable. However, according to a study by Woodland (1979) the normality assumption yields results very close to those obtained with a Dirichlet distribution. He concludes (p. 302) that "...while the Normal model may not be a theoretically appropriate specification for share equations, it may, for a large number of data sets, yield valid results."

The vector $\boldsymbol{s}^+_{t\cdot}$ of predicted cost shares can be expressed in matrix terms as follows. The exogenous variables can be compiled into the $T \times 3$ matrix $\mathbf{W}$ whose tth row is $\boldsymbol{w}_{t\cdot} = (w_{t1}, w_{t2}, w_{t3})$ and the $T \times 5$ matrix $\mathbf{Z}$ whose tth row is

$$\boldsymbol{z}_{t\cdot} \equiv (1, \ln w_{t1}, \ln w_{t2}, \ln w_{t3}, \tfrac{t}{2}), \quad t = 1,\ldots,T.$$

Both $\mathbf{W}$ and $\mathbf{Z}$ are assumed to have full column rank.

There are two coefficient matrices whose elements are functions of the parameters to be estimated. The first one, denoted $\mathbf{B}$, which is 3×5 and dependent on the subvector $\boldsymbol{\kappa}$ of $\boldsymbol{\theta}$, is given by :

$$\mathbf{B}(\boldsymbol{\kappa}) = \begin{bmatrix} b_{11} & \gamma_{11}/\lambda & \gamma_{12}/\lambda & -(\gamma_{11}+\gamma_{12})/\lambda & \tau_1/\lambda \\ b_{21} & \gamma_{12}/\lambda & \gamma_{22}/\lambda & -(\gamma_{12}+\gamma_{22})/\lambda & \tau_2/\lambda \\ b_{31} & -(\gamma_{11}+\gamma_{12}) & -(\gamma_{12}+\gamma_{22}) & (\gamma_{11}+2\gamma_{12}+\gamma_{22}) & -(\tau_1+\tau_2) \end{bmatrix} \tag{20}$$

where

$$b_{11} = [\alpha_1 + (\gamma_{11}+\gamma_{12})\cdot\ln\lambda]/\lambda$$
$$b_{21} = [\alpha_2 + (\gamma_{12}+\gamma_{22})\cdot\ln\lambda]/\lambda$$
$$b_{31} = (1-\alpha_1-\alpha_2) - (\gamma_{11}+2\gamma_{12}+\gamma_{22})\cdot\ln\lambda\,,$$

and

$$\lambda = \exp(\beta)\,.$$

The second coefficient matrix, $\mathbf{M}$, is 3×3 diagonal. Its diagonal elements are functions of the 3×1 vector $\boldsymbol{\nu}$ according to

$$\mathbf{M} = \mathbf{M}[\boldsymbol{\mu}(\boldsymbol{\nu})] = \mathrm{diag}[\boldsymbol{\mu}(\boldsymbol{\nu})]$$
$$= \mathrm{diag}\{[\exp(\nu_1), \exp(\nu_2), \exp(\nu_3)]'\}.\text{[17]} \tag{21}$$

[17] For a definition of the diag operator, see Appendix A, eq. (A.17).

Using (11) and (12) and the above notation one can express $s^+_{t\cdot}$ as follows

$$s^+_{t\cdot} = \frac{(g^\star_t \cdot \mathbf{J}\mathbf{B}z'_{t\cdot} + \mathbf{J}\mathbf{M}w'_{t\cdot})'}{g^\star_t \cdot \mathbf{1}'_3\mathbf{B}z'_{t\cdot} + \mathbf{1}'_3\mathbf{M}w'_{t\cdot}} \tag{22}$$

where $\mathbf{1}_3$ is the 3×1 unit vector and $\mathbf{J}$ is a selection matrix, defined according to

$$\mathbf{J} \equiv (\mathbf{I}_2 : \mathbf{0}_2) \,. \tag{23}$$

Premultiplication of $\mathbf{B}$ and $\mathbf{M}$ by $\mathbf{J}$ thus has the effect of selecting the 2 first rows of these matrices. The scalar $g^\star_t$, finally, can be written

$$\begin{aligned} g^\star_t &= g(w_{t\cdot}, t\,;\,\kappa) \\ &= \exp\Big[[\ln(w^\star_{t\cdot})]\alpha + \tfrac{1}{2}[\ln(w^\star_{t\cdot})]\Gamma\,[\ln(w^\star_{t\cdot})]' + \tfrac{1}{2}[t\cdot\ln(w^\star_{t\cdot})]\tau\Big] \end{aligned} \tag{24}$$

where

$$\ln(w^\star_{t\cdot}) \equiv [\ln(\lambda w_{t1}), \ln(\lambda w_{t2}), \ln(w_{t3})] \,.$$ [18]

The scalar λ, the 3×1 vectors α and τ, and the 3×3 matrix Γ are functions of θ_1, θ_2, θ_3, and θ_4 according to

$$\alpha = (\theta'_1, \alpha_n)' \,, \qquad \alpha_n = 1 - \theta'_1\mathbf{1}_2 \,, \tag{25a}$$

$$\tau = (\theta'_3, \tau_n)' \,, \qquad \tau_n = -\,\theta'_3\mathbf{1}_2 \,, \tag{25b}$$

$$\lambda = \lambda(\theta_4) = \exp(\theta_4) \,, \tag{25c}$$

$$\Gamma = (\gamma_{ij}) = \Gamma(\theta_2)\,; \quad \Gamma = \Gamma', \quad \Gamma\mathbf{1}_3 = (\mathbf{1}'_3\Gamma)' = \mathbf{0}_3. \tag{25d}$$

In matrix form, comprising all T observations, the system is given by

$$\mathbf{R}' = \mathbf{S}^{+\prime} + \mathbf{U}' = \mathbf{J}(\mathbf{B}\mathbf{Z}'\mathbf{G} + \mathbf{M}\mathbf{W}')\mathbf{D}^{-1} + \mathbf{U}' \,, \tag{26}$$

---

[18] The function $g(w_{t\cdot}, t\,;\,\kappa)$ and the function denoted $g(w^\star, t)$ in Section 2 are one and the same. The former notation is better suited in the present context as it makes an explicit distinction between parameters and variables.

where $\mathbf{R}' = (\boldsymbol{r}'_{1\cdot}, \boldsymbol{r}'_{2\cdot}, \ldots, \boldsymbol{r}'_{T\cdot})$ and $\mathbf{S}^{+\prime}$ and $\mathbf{U}'$ are defined analogously,

$$\mathrm{G} = \mathrm{diag}(\boldsymbol{g}^{\star})] = \mathrm{diag}[(g_1^{\star}, \ldots, g_T^{\star})'] \,, \tag{27}$$

and

$$\mathrm{D} = \mathrm{diag}[(1_3'\mathrm{BZ}'\mathrm{G} + 1_3'\mathrm{MW}')'] \,. \tag{28}$$

Notice that, in contrast to the matrices **Z**, **W**, **B**, and **M** whose elements are either exogenous variables or (functions of) parameters, the elements of the matrix G are functions of *both* the exogenous variables and the parameter; cf. (24).

5. The likelihood function and its first order derivatives

From (19) it can be seen that, in statistical terms, the system of input cost shares constitutes a multivariate regression model. In spite of the strong non-linearities it involves the model is actually not overly complicated to analyze. The main reason for this is that the non-linearities are confined to the mapping from the vector $\boldsymbol{\theta}$ of unrestricted parameters to the vector $\boldsymbol{s}^{+}_{t\cdot}$ of predicted cost shares. This property implies that several well known results concerning maximum likelihood estimation of multivariate regression models apply. In particular, general non-linear models of the form (19) have been considered by Berndt et al. (1974). In principle, the problem here is to extend their analysis so as to explicitly account for the particular form of the non-linear functions connecting $\boldsymbol{\theta}$ and $\boldsymbol{s}^{+}_{t\cdot}$.

Another feature which facilitates the analysis is the fact that the restrictions (3), (7), and (8) have been used in the derivation of the predicted cost shares. This means that the problem of finding the parameter estimates can be formulated as an unconstrained optimization problem.

It should be noted that neither of two properties just described have anything to do with the number of inputs or the constraint (7c). Accordingly, they will apply also in more general contexts when $n > 3$ and the degree of allocative inefficiency is allowed to vary between different pairs of inputs.

While Berndt et al. (op. cit.) work with the concentrated log-likelihood function, the unconcentrated function will be used here because it provides a more natural starting point for the analysis of the information matrix in the next section. By considering the unconcentrated log-likelihood function one can also avoid a small inconsistency in the results of Berndt et al., concerning their expression for the first order derivatives of the log-likelihood function. Whereas the expression that they provide always can be obtained from the unconcentrated log-likelihood function, it is only asymptotically true for the concentrated log-likelihood function, from which Berndt et al. claim that it has been derived.

As mentioned above, the new problem to be addressed here relates to the specific form of the non-linear mapping from the vector of unrestricted parameters to the vector of predicted cost shares. Accordingly, this section commences with a lemma concerning the first order partial derivatives of this mapping. By means of the lemma a theorem is then proved, which extends the results of Berndt et al. to the present model.

In the calculations, references will frequently be made to equations in Appendix A and Appendix B. The former have the prefix A and the latter the prefix B. Concerning the results on differentiation in Appendix B, the definition (B.1) should be noted, according to which the derivative of the $m \times 1$ vector $\boldsymbol{y}$ with respect to the $n \times 1$ vector $\boldsymbol{x}$ is the $m \times n$ matrix $\partial \boldsymbol{y} / \partial \boldsymbol{x}$ with typical element $\partial y_i / \partial x_j$, i and j being row and column indices, respectively.[19]

LEMMA 1. *The partial derivative of* $(\mathbf{S}^{+\prime})^c = [\mathbf{J}(\mathbf{BZ'G} + \mathbf{MW'})\mathbf{D}^{-1}]^c$ *with respect to* $\boldsymbol{\theta} = (\boldsymbol{\kappa}', \boldsymbol{\nu}')'$ *is given by*

$$\frac{\partial(\mathbf{S}^{+\prime})^c}{\partial\boldsymbol{\theta}} = (\mathbf{D}^{-1} \otimes \mathbf{I}_2)\, \mathbf{A}' \left[(\mathbf{GZ} \otimes \mathbf{I}_3)\, \frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\theta}} + (\mathbf{W} \otimes \mathbf{I}_3)\, \frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\theta}} + (\mathbf{I}_T \otimes \mathbf{BZ'})\, \frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\theta}}\right],$$

where

$$\mathbf{A} \equiv (\mathbf{I}_T \otimes \mathbf{J}') - (\mathbf{I}_T \otimes \mathbf{1}_3)\boldsymbol{\Psi}_T(\mathbf{I}_T \otimes \mathbf{S}^+) \tag{29}$$

and $\boldsymbol{\Psi}_T$ *is the* $T \times T^2$ *basis matrix for diagonality defined in Magnus* (1988, *p.* 109) *and in* (A.20).

Further, the partial derivatives $\partial\mathbf{B}^c/\partial\boldsymbol{\theta}$, $\partial\mathbf{M}^c/\partial\boldsymbol{\theta}$, *and* $\partial\mathbf{G}^c/\partial\boldsymbol{\theta}$ *can be expanded according to*

$$\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\theta}} = \left[\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_{5\cdot3,3}\right]$$

$$\frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\theta}} = \left[\mathbf{0}_{3^2,8} \,\vdots\, \boldsymbol{\Psi}_3'\mathbf{M}\right]$$

$$\frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\theta}} = \left[\boldsymbol{\Psi}_T' \frac{\partial\boldsymbol{g}^\star}{\partial\boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_{T^2,3}\right].$$

where $\vdots$ *denotes partitioned matrix, i.e.* $\left[\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_{5\cdot3,3}\right]$ *is partitioned into the* $5\cdot3 \times 8$ *matrix* $\partial\mathbf{B}^c/\partial\boldsymbol{\kappa}$ *and the* $5\cdot3 \times 3$ *zero matrix.*

PROOF. See Appendix C.

[19] While this definition, due to Pollock (1979), is at variance with much of the early work on matrix differential calculus it is now becoming widely accepted. Recently, the same definition has also been strongly advocated by Magnus and Neudecker (1988). However, they denote the matrix of partial derivatives by $\partial\boldsymbol{y}/\partial\boldsymbol{x}'$.

By means of Lemma 1 one can prove the following theorem.

THEOREM 1. *For the model specified in Section 4 the log-likelihood function is*

$$L(\boldsymbol{\theta},\boldsymbol{\Omega}) = \text{const.} - \tfrac{\mathrm{T}}{2}\cdot\ln|\boldsymbol{\Omega}| - \tfrac{1}{2}\cdot\mathrm{tr}[(\mathbf{R}-\mathbf{S}^{+})'(\mathbf{R}-\mathbf{S}^{+})\boldsymbol{\Omega}^{-1}] . \tag{30}$$

The first order derivative of $L(\boldsymbol{\theta},\boldsymbol{\Omega})$ *with respect to* $\boldsymbol{\Omega}$ *is given by*

$$\frac{\partial L(\boldsymbol{\theta},\boldsymbol{\Omega})}{\partial \boldsymbol{\Omega}^{c}} = \tfrac{1}{2}\Big[\boldsymbol{\Omega}^{-1}(\mathbf{R}-\mathbf{S}^{+})'(\mathbf{R}-\mathbf{S}^{+})\boldsymbol{\Omega}^{-1}\Big]^{c'} - \tfrac{\mathrm{T}}{2}\cdot\boldsymbol{\Omega}^{-1c'} . \tag{31}$$

Further, two equivalent expressions for the first order derivative of $L(\boldsymbol{\theta},\boldsymbol{\Omega})$ *with respect to* $\boldsymbol{\theta} = (\boldsymbol{\kappa}' , \boldsymbol{\nu}')'$ *are*

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta},\boldsymbol{\Omega})}{\partial \boldsymbol{\theta}} = {} & [(\mathbf{F}-\mathbf{v}\mathbf{1}_3')']^{c'} \Big[(\mathbf{GZ}\otimes\mathbf{I}_3)\Big[\frac{\partial \mathbf{B}^{c}}{\partial \boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_{5\cdot 3,3}\Big] \\ & + (\mathbf{W}\otimes\mathbf{I}_3)\Big[\mathbf{0}_{3^2,8} \,\vdots\, \boldsymbol{\Psi}_3'\mathbf{M}\Big] + (\mathbf{I}_{\mathrm{T}}\otimes\mathbf{BZ}')\Big[\boldsymbol{\Psi}_{\mathrm{T}}'\frac{\partial \boldsymbol{g}^{\star}}{\partial \boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_{\mathrm{T}^2,3}\Big]\Big]\end{aligned} \tag{32}$$

and

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta},\boldsymbol{\Omega})}{\partial \boldsymbol{\theta}} = {} & \Big[[\mathbf{F}-\mathbf{v}\mathbf{1}_3')'\mathbf{GZ}]^{c'}\frac{\partial \mathbf{B}^{c}}{\partial \boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_3'\Big] + \Big[\mathbf{0}_8' \,\vdots\, [(\mathbf{F}-\mathbf{v}\mathbf{1}_3')'\mathbf{W}]^{d'}\mathbf{M}\Big] \\ & + \Big[[\mathbf{ZB}'(\mathbf{F}-\mathbf{v}\mathbf{1}_3')']^{d'}\frac{\partial \boldsymbol{g}^{\star}}{\partial \boldsymbol{\kappa}} \,\vdots\, \mathbf{0}_3'\Big] ,\end{aligned} \tag{33}$$

where $\mathbf{X}^{d}$ *denotes the column vector corresponding to diagonal elements of the square matrix* $\mathbf{X}$ [*cf.* (A.18)] *and*

$$\mathbf{F} \equiv \mathbf{D}^{-1}(\mathbf{R}-\mathbf{S}^{+})\boldsymbol{\Omega}^{-1}\mathbf{J} , \tag{34}$$

$$\mathbf{v} \equiv [\mathbf{S}^{+}\boldsymbol{\Omega}^{-1}(\mathbf{R}-\mathbf{S}^{+})'\mathbf{D}^{-1}]^{d} , \tag{35a}$$

$$\equiv \boldsymbol{\Psi}_{\mathrm{T}}\, [\mathbf{S}^{+}\boldsymbol{\Omega}^{-1}(\mathbf{R}-\mathbf{S}^{+})'\mathbf{D}^{-1}]^{c} . \tag{35b}$$

PROOF. The likelihood function can be found in, e.g., Pollock (1979, p. 236).

To prove (31), note that, by the chain rule (B.5)

$$\frac{\partial L(\boldsymbol{\theta},\Omega)}{\partial \Omega^{c}} = -\frac{T}{2}\cdot\frac{\partial \ln|\Omega|}{\partial \Omega^{c}} - \frac{1}{2}\cdot\frac{\partial \mathrm{tr}[(\mathbf{R}-\mathbf{S}^{+})'(\mathbf{R}-\mathbf{S}^{+})\Omega^{-1}]}{\partial \Omega^{-1\,c}}\,\frac{\partial \Omega^{-1c}}{\partial \Omega^{c}}\,.$$

Application of (B8), (B.9), (B.7) and, finally, (A.3a) to this expression yields (31). Concerning (32) and (33), application of the chain rule gives

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta},\Omega)}{\partial \boldsymbol{\theta}} &= -\frac{1}{2}\cdot\frac{\partial \mathrm{tr}[\cdot]}{\partial(\mathbf{R}-\mathbf{S}^{+})^{c}}\,\frac{\partial(\mathbf{R}-\mathbf{S}^{+})^{c}}{\partial(\mathbf{S}^{+})^{c}}\,\frac{\partial(\mathbf{S}^{+})^{c}}{\partial(\mathbf{S}^{+\prime})^{c}}\,\frac{\partial(\mathbf{S}^{+\prime})^{c}}{\partial \boldsymbol{\theta}} \\
&= -\frac{1}{2}\cdot[2(\mathbf{R}-\mathbf{S}^{+})^{c\prime}(\Omega^{-1}\otimes \mathbf{I}_{T})](-\mathbf{I}_{2\cdot T})\mathbf{K}_{2T}\,\frac{\partial(\mathbf{S}^{+\prime})^{c}}{\partial \boldsymbol{\theta}} \\
&= [(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{I}_{T}\otimes \Omega^{-1})\,\frac{\partial(\mathbf{S}^{+\prime})^{c}}{\partial \boldsymbol{\theta}}\,, \qquad (36)
\end{aligned}$$

where the second equality follows from (B.10) and (B.4) and the third equality is obtained by means of (A.15), (A.3b), (A.5), and (A.3a). Like (31), this result, which does not take the specific form of the mapping from $\boldsymbol{\theta}$ to $\mathbf{S}^{+}$ into account, is just a variant of a result that is well known in the literature. An equivalent expression can, e.g., be found in Berndt et al. (1974, p. 663).

To proceed, use the first part of Lemma 1 in (36) and rearrange, to get

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta},\Omega)}{\partial \boldsymbol{\theta}} &= [(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{D}^{-1}\otimes \Omega^{-1})\Big[(\mathbf{GZ}\otimes \mathbf{J})\,\frac{\partial \mathbf{B}^{c}}{\partial \boldsymbol{\theta}} \\
&\quad + (\mathbf{W}\otimes \mathbf{J})\,\frac{\partial \mathbf{M}^{c}}{\partial \boldsymbol{\theta}} + (\mathbf{I}_{T}\otimes \mathbf{JBZ}')\,\frac{\partial \mathbf{G}^{c}}{\partial \boldsymbol{\theta}}\Big] \\
&\quad - [(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{D}^{-1}\otimes \Omega^{-1}\mathbf{S}^{+\prime})\,\Psi_{T}'\Big[(\mathbf{GZ}\otimes \mathbf{1}_{3}')\,\frac{\partial \mathbf{B}^{c}}{\partial \boldsymbol{\theta}} \\
&\quad + (\mathbf{W}\otimes \mathbf{1}_{3}')\,\frac{\partial \mathbf{M}^{c}}{\partial \boldsymbol{\theta}} + (\mathbf{I}_{T}\otimes \mathbf{1}_{3}'\mathbf{BZ}')\,\frac{\partial \mathbf{G}^{c}}{\partial \boldsymbol{\theta}}\Big]\,. \qquad (36')
\end{aligned}$$

Note that, by (A.3a) and (A.14)

$$[(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{D}^{-1}\otimes \Omega^{-1}) = [\mathbf{D}^{-1}(\mathbf{R}-\mathbf{S}^{+})\Omega^{-1}]^{r} \qquad (37)$$

and

$$(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{D}^{-1}\otimes\Omega^{-1}\mathbf{S}^{+\prime})\,\Psi_{\mathrm{T}}' = [\mathbf{D}^{-1}(\mathbf{R}-\mathbf{S}^{+})\Omega^{-1}\mathbf{S}^{+\prime}]^{r}\Psi_{\mathrm{T}}' . \tag{38}$$

Further, by means of (A.3a), (A.14), (A.21) and (A.2'), (38) can be rewritten according to

$$\begin{aligned}
[(\mathbf{R}-\mathbf{S}^{+})']^{c\prime}\,(\mathbf{D}^{-1}\otimes\Omega^{-1}\mathbf{S}^{+\prime})\,\Psi_{\mathrm{T}}' &= \{\Psi_{\mathrm{T}}\,[\mathbf{D}^{-1}(\mathbf{R}-\mathbf{S}^{+})\Omega^{-1}\mathbf{S}^{+\prime}]^{r\prime}\}' \\
&= \{\Psi_{\mathrm{T}}\,[\mathbf{S}^{+}\Omega^{-1}(\mathbf{R}-\mathbf{S}^{+})'\mathbf{D}^{-1}]^{c}\}' = [\mathbf{S}^{+}\Omega^{-1}(\mathbf{R}-\mathbf{S}^{+})'\mathbf{D}^{-1}]^{d\prime} \\
&= \{[\mathbf{S}^{+}\Omega^{-1}(\mathbf{R}-\mathbf{S}^{+})'\mathbf{D}^{-1}]^{d}\}^{r} .
\end{aligned} \tag{38'}$$

By inserting (37) and (38') in (36'), using (34) and (35a), and (A.14) one gets

$$\frac{\partial L(\boldsymbol{\theta},\Omega)}{\partial\boldsymbol{\theta}} = [\mathbf{ZG}(\mathbf{F}-\mathbf{v}\mathbf{1}_3')]^{r}\,\frac{\partial\mathbf{B}^{c}}{\partial\boldsymbol{\kappa}} + [\mathbf{W}'(\mathbf{F}-\mathbf{v}\mathbf{1}_3')]^{r}\,\frac{\partial\mathbf{M}^{c}}{\partial\boldsymbol{\theta}} + [(\mathbf{F}-\mathbf{v}\mathbf{1}_3')\mathbf{BZ}']^{r}\,\frac{\partial\mathbf{G}^{c}}{\partial\boldsymbol{\theta}} .$$

Given this result, (32) follows from application of (A.14) and (A.3a), and the second part of Lemma 1. The alternative form (33) is obtained by inserting the second part of Lemma 1 and applying (A.3a) and (A.21). Q.E.D.

The reason for giving the two equivalent expressions for the gradient vector is that while the form (33) is more convenient for evaluating the first order derivatives, the form (32) is better suited for the calculation of the information matrix, which is the topic of the next section.

6. The information matrix

The principal reason for considering the information matrix is that this matrix can be used to check whether the model's parameters are identified. As shown by Rothenberg (1971) the parameters are (locally) identified if, and only if, the information matrix has full rank, i.e. is non-singular. In addition, given that the parameters are identified, the asymptotic covariance matrix of the parameter estimates can be estimated by means of (the sample analogue of) the information matrix.[20] Finally, estimates of the information matrix can be useful in the construction of numerical algorithms for the generation of the parameter estimates themselves.[21]

According to a well-known property of the likelihood function, the information matrix $\mathcal{I} = \mathcal{I}(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)$, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\Omega}_0$ denote the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, can be defined in two equivalent ways. According to the most common definition it is given by the negative of the expected value of the Hessian matrix of the log-likelihood function, evaluated at $(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)$. However, the information matrix may also be defined as expected value of the outer product of the first order derivatives of the log-likelihood function, again evaluated at $(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)$. [22] In the following, both definitions will be used.

With respect to the particular model at hand, the information matrix is block-diagonal, due to the fact that there are no restrictions on the covariance matrix of the disturbances, $\boldsymbol{\Omega}$; cf. Magnus and Neudecker (1988, Theorem 7, p. 326). The information matrix can thus be written

$$\mathcal{I} = \mathcal{I}(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0) = \begin{bmatrix} \mathcal{I}(\boldsymbol{\theta}_0) & \mathbf{0}_{11,4} \\ \mathbf{0}_{4,11} & \mathcal{I}(\boldsymbol{\Omega}_0) \end{bmatrix} \tag{39}$$

where $\mathbf{0}_{11,4}$ denotes the 11×4 zero matrix,

$$\mathcal{I}(\boldsymbol{\theta}_0) = -E\left[\frac{\partial[\partial L(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)/\partial\boldsymbol{\theta}_0]'}{\partial\boldsymbol{\theta}_0}\right] = E\left[\left[\frac{\partial L(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)}{\partial\boldsymbol{\theta}}\right]'\left[\frac{\partial L(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)}{\partial\boldsymbol{\theta}}\right]\right]$$

[20] See, e.g., Magnus and Neudecker (1988, p. 314).

[21] Two examples are the method of scoring and the method proposed by Berndt et al. (1974). An elementary discussion of both of these can be found in Maddala (1977, pp. 176 – 179.).

[22] Berndt et al. (1974) were the first to exploit this property in an econometric context. For a formal proof of the equivalence, see, e.g., Pollock (1979, p. 339).

and

$$\mathcal{I}(\Omega_0) = -E\left[\frac{\partial\left[\partial L(\theta_0,\Omega_0)/\partial\Omega^c\right]'}{\partial\Omega^c}\right] = E\left[\left[\frac{\partial L(\theta_0,\Omega_0)}{\partial\Omega^c}\right]'\left[\frac{\partial L(\theta_0,\Omega_0)}{\partial\Omega^c}\right]\right].$$

Clearly, if $\mathcal{I}(\theta_0)$ and $\mathcal{I}(\Omega_0)$ are of full rank, then $\mathcal{I}(\theta_0,\Omega_0)$ must have full rank, too. Concerning the present model, it is a simple matter to derive the submatrix $\mathcal{I}(\Omega_0)$ and to show that it has full rank. These results, which are well known, are stated in in the following lemma.

LEMMA 2. *The submatrix* $\mathcal{I}(\Omega_0)$ *of* $\mathcal{I}(\theta_0,\Omega_0)$ *is given by*

$$\mathcal{I}(\Omega_0) = \frac{T}{2}(\Omega^{-1}\otimes\Omega^{-1})$$

and has full rank, i.e. $\text{rank}[\mathcal{I}(\Omega_0)] = 4$.

PROOF. See Appendix C.

Given Lemma 2, proving the full rank of $\mathcal{I}(\theta_0,\Omega_0)$ amounts to proving the full rank of $\mathcal{I}(\theta_0)$. One way to establish the rank of $\mathcal{I}(\theta_0)$ is suggested in the following proposition.

PROPOSITION 1. *The submatrix* $\mathcal{I}(\theta_0)$ *of* $\mathcal{I}(\theta_0,\Omega_0)$ *satisfies the following two conditions.*

(i) *It can partitioned according to*

$$\mathcal{I}(\theta_0) = \mathcal{I}(\kappa_0',\nu_0')' = \begin{bmatrix}\Upsilon_{11} & \Upsilon_{12}\\ \Upsilon_{12}' & \Upsilon_{22}\end{bmatrix}, \tag{40}$$

where Υ_{11} , Υ_{12} , and Υ_{22} *are defined according to*

$$\Upsilon_{11} \equiv E\left[\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\kappa}}\right]'\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\kappa}}\right]\right],$$

$$\Upsilon_{12} \equiv E\left[\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\kappa}}\right]'\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\nu}}\right]\right],$$

$$\Upsilon_{22} \equiv E\left[\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\nu}}\right]'\left[\frac{\partial L(\boldsymbol{\theta}_0,\Omega_0)}{\partial \boldsymbol{\nu}}\right]\right],$$

and equal to

$$\Upsilon_{11} = \left[\left[\frac{\partial \mathrm{B}^c}{\partial \boldsymbol{\kappa}}\right]'(\mathrm{Z'G} \otimes \mathrm{I}_n) + \left[\frac{\partial \boldsymbol{g}^\star}{\partial \boldsymbol{\kappa}}\right]'\Psi_T(\mathrm{I}_T \otimes \mathrm{ZB'})\right]\mathrm{A}(\mathrm{D}^{-2} \otimes \Omega^{-1})\mathrm{A}'$$
$$\times \left[(\mathrm{GZ} \otimes \mathrm{I}_n)\left[\frac{\partial \mathrm{B}^c}{\partial \boldsymbol{\kappa}}\right] + (\mathrm{I}_T \otimes \mathrm{BZ'})\Psi_T'\left[\frac{\partial \boldsymbol{g}^\star}{\partial \boldsymbol{\kappa}}\right]\right], \tag{41}$$

$$\Upsilon_{12} = \left[\left[\frac{\partial \mathrm{B}^c}{\partial \boldsymbol{\kappa}}\right]'(\mathrm{Z'G} \otimes \mathrm{I}_n) + \left[\frac{\partial \boldsymbol{g}^\star}{\partial \boldsymbol{\kappa}}\right]'\Psi_T(\mathrm{I}_T \otimes \mathrm{ZB'})\right]\mathrm{A}(\mathrm{D}^{-2} \otimes \Omega^{-1})\mathrm{A}'$$
$$\times (\mathrm{W} \otimes \mathrm{I}_n)\Psi_n'\mathrm{M}, \tag{42}$$

$$\Upsilon_{22} = \mathrm{M}\Psi_n(\mathrm{W}' \otimes \mathrm{I}_n)\mathrm{A}(\mathrm{D}^{-2} \otimes \Omega^{-1})\mathrm{A}'(\mathrm{W} \otimes \mathrm{I}_n)\Psi_n'\mathrm{M}, \tag{43}$$

the matrix A *being defined by* (29).

(ii) *It has full rank if* Υ_{11} , Υ_{22} , *and* $(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}')$ *all have full ranks.*

PROOF. To prove part (i), first take the expectation of the outer product of the expression (36), which was derived in the proof of Theorem (1) :

$$\mathcal{I}(\boldsymbol{\theta}_0) = E\left[\left[\frac{\partial(\mathrm{S}^{+\prime})^c}{\partial \boldsymbol{\theta}}\right]'(\mathrm{I}_T \otimes \Omega^{-1})\,(\mathrm{R}-\mathrm{S}^+)'^c(\mathrm{R}-\mathrm{S}^+)'^{c\prime}(\mathrm{I}_T \otimes \Omega^{-1})\,\frac{\partial(\mathrm{S}^{+\prime})^c}{\partial \boldsymbol{\theta}}\right]$$
$$= \left[\frac{\partial(\mathrm{S}^{+\prime})^c}{\partial \boldsymbol{\theta}}\right]'(\mathrm{I}_T \otimes \Omega^{-1})\,\frac{\partial(\mathrm{S}^{+\prime})^c}{\partial \boldsymbol{\theta}}. \tag{44}$$

The second equality follows since $E\left[(\mathrm{R}-\mathrm{S}^+)'^c(\mathrm{R}-\mathrm{S}^+)'^{c\prime}\right] = (\mathrm{I}_T \otimes \Omega)$. The result (44) is not new; see, e.g., Berndt et al. (1974, p. 664) for an equivalent expression.

The next step is to expand (44) by means of Lemma 1. The the first part of Lemma 1 yields,

$$\mathcal{I}(\boldsymbol{\theta}_o) = \left[\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \left[\frac{\partial \mathbf{G}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}') + \left[\frac{\partial \mathbf{M}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{W}' \otimes \mathbf{I}_3)\right]$$

$$\times\ \mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}' \times \tag{45}$$

$$\left[\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \left[\frac{\partial \mathbf{G}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}') + \left[\frac{\partial \mathbf{M}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{W}' \otimes \mathbf{I}_3)\right]',$$

where the RHS is implicitly taken to be evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\Omega}_0)$. Finally, by the second part of Lemma 1,

$$\left[\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \left[\frac{\partial \mathbf{G}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}') + \left[\frac{\partial \mathbf{M}^c}{\partial \boldsymbol{\theta}}\right]'(\mathbf{W}' \otimes \mathbf{I}_3)\right]$$

$$= \begin{bmatrix} \left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\kappa}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) \\ \mathbf{0}_{3,T\cdot 3} \end{bmatrix} + \begin{bmatrix} \left[\frac{\partial \boldsymbol{q}^{\star}}{\partial \boldsymbol{\kappa}}\right]'\boldsymbol{\Psi}_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}') \\ \mathbf{0}_{3,T\cdot 3} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{8,T\cdot 3} \\ \mathbf{M}\boldsymbol{\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3) \end{bmatrix}.$$

Inserting the RHS of this last expression in (45) one obtains the desired result, after a number of tedious but straightforward calculations.

The second part of the Lemma is an immediate consequence of well known results concerning inverses of partitioned matrices. If $\boldsymbol{\Upsilon}_{11}$, $\boldsymbol{\Upsilon}_{22}$, and $(\boldsymbol{\Upsilon}_{11} - \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}_{12}')$ all have full rank then the inverse of $\mathcal{I}(\boldsymbol{\theta}_o)$ exists, which implies that it must have full rank: see, e.g., Johnston (1984, p. 135). Q.E.D.

To show that $\boldsymbol{\Upsilon}_{11}$, $\boldsymbol{\Upsilon}_{22}$, and $(\boldsymbol{\Upsilon}_{11} - \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}_{12}')$ have full rank it is necessary to consider these matrices in detail. To this end, the following two lemmas, which characterize the matrices $\mathbf{A}$ and $\mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}'$, respectively, will be useful. The first lemma demonstrates, that in spite of its complicated appearance, the matrix $\mathbf{A}$ has a quite simple block-diagonal structure.

LEMMA 3. *The* $T\cdot 3 \times T\cdot 2$ *matrix* $\mathbf{A} \equiv (\mathbf{I}_T \otimes \mathbf{J}') - (\mathbf{I}_T \otimes \mathbf{1}_3)\boldsymbol{\Psi}_T((\mathbf{I}_T \otimes \mathbf{S}^+)$ *has the following properties :*

(i) *It can be written in block-diagonal form according to*

$$\mathbf{A} = \begin{bmatrix} \mathbf{J}' - \mathbf{1}_3\boldsymbol{s}_{1\cdot}^{+} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}' - \mathbf{1}_3\boldsymbol{s}_{2\cdot}^{+} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}' - \mathbf{1}_3\boldsymbol{s}_{T\cdot}^{+} \end{bmatrix},$$

where $\boldsymbol{s}_{t\cdot}^{+}$, $t = 1,\ldots,T$, *denotes the t'th row of the matrix* $\mathbf{S}^+$ *given by* (26) *and the zero matrices are* 3×2.

(ii) *It has full column rank, i.e.* $\text{rank}(\mathbf{A}) = T\cdot 2$.

PROOF. See Appendix C.

The next lemma provides a decomposition of $\mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}'$ and establishes its rank.

LEMMA 4. The $T\cdot 3 \times T\cdot 3$ *matrix* $\mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}'$ *has the following properties.*

(i) *It can be decomposed according to*

$$\mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}' = [\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})][\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})]'$$

where $\mathbf{P}$ *is a non-singular* 2×2 *matrix.*

(ii) $\text{Rank}[\mathbf{A}(\mathbf{D}^{-2} \otimes \boldsymbol{\Omega}^{-1})\mathbf{A}'] = \text{rank}[\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})] = \text{rank}(\mathbf{A})$.

PROOF. See Appendix C.

Since $\boldsymbol{\Upsilon}_{22}$ is the less complicated of the three matrices whose rank are to be determined, it will be considered first. The following lemma gives a detailed description of a product matrix involved in Υ_{22} and in the theorem immediately following it is proved that Υ_{22} has full rank.

LEMMA 5. *The* $3 \times T{\cdot}2$ *matrix* $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}$ *involved in* Υ_{22} *can be partitioned into* T *submatrices of dimension* 3×2 *according to* :

$$\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A} = [\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot})(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}^+_{t\cdot}) \ , \quad t = 1,\dots,T\,]$$

where $\boldsymbol{w}_{t\cdot}$, $t = 1,\dots,T,$ *denotes the* t*'th row of* $\mathbf{W}$, $\boldsymbol{\mu}$ *is implicitly defined by the equality* $\mathbf{M} = \mathrm{diag}(\boldsymbol{\mu})$, *and* $\odot$ *denotes the Hadamard product, i.e.* $\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot} = (\mu_1 w_{t1}, \dots, \mu_n w_{tn})'$.

PROOF. See Appendix C.

THEOREM 2. *The* 3×3 *matrix* Υ_{22} *given by* (43) *has full rank.*

PROOF. By (43), Lemma 4 and the results concerning ranks that were exploited in the proof of Lemma 4 :

$$\begin{aligned}
\mathrm{rank}(\Upsilon_{22}) &= \mathrm{rank}\{[\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})][\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})]'\} \\
&= \mathrm{rank}[\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})] \\
&= \mathrm{rank}[\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}] \ .
\end{aligned}$$

Since $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}$ is $3 \times T{\cdot}2$ the task is to prove that $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}$ has full row rank. To this end, partition $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathbf{I}_3)\mathbf{A}$ in accordance with Lemma 5, and note that

$$\mathrm{rank}[\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot})(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}^+_{t\cdot})] = \mathrm{Rank}(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}^+_{t\cdot}) = 2 \ , \quad t = 1,\dots,T. \tag{46}$$

The first equality follows from the non-singularity of $\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot})$ and the second from Lemma 3. By (46), the rank of $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathrm{I}_3)\mathbf{A}$ must at least be equal to that of its submatrices $[\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot})(\mathbf{J}' - 1_3\boldsymbol{s}^+_{t\cdot})]$, $t = 1,\dots,T$. Thus, given (46), the theorem is equivalent to the claim that the row rank of $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathrm{I}_3)\mathbf{A}$ *exceeds* that of its T submatrices. This will be verified by showing that the rank of $\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathrm{I}_3)\mathbf{A}$ cannot be equal to the rank of its submatrices unless the predicted factor proportions are all constant over time.

Consider the t'th submatrix. As this submatrix is 3×2 and its rank is 2 there exists one linear combination of its first 2 rows which is equal to its last row. Noting that by the definition (23) of $\mathbf{J}$ the tth submatrix can be partitioned according to

$$\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}'_{t\cdot})(\mathbf{J}' - 1_3\boldsymbol{s}^+_{t\cdot}) = \begin{bmatrix} \mathrm{diag}[(\mu_1 w_{t1}, \mu_2 w_{t2})'](\mathrm{I}_2 - 1_2\boldsymbol{s}^+_{t\cdot}) \\ -\mu_3 w_{t3}\boldsymbol{s}^+_{t\cdot} \end{bmatrix}$$

this means that for $t = 1,\dots,T$, there is a 2×1 vector $\boldsymbol{a}_t \in \mathbb{R}^2$ such that

$$\boldsymbol{a}'_t\left[\mathrm{diag}[(\mu_1 w_{t1}, \mu_2 w_{t2})'](\mathrm{I}_2 - 1_2\boldsymbol{s}^+_{t\cdot})\right] = -\mu_3 w_{t3}\boldsymbol{s}^+_{t\cdot}\ .$$

Since

$$(\mathrm{I}_2 - 1_2\boldsymbol{s}^+_{t\cdot})^{-1} = (\mathrm{I}_2 + \delta_t^{-1}\cdot 1_2\boldsymbol{s}^+_{t\cdot}) \tag{47a}$$

where

$$\delta_t \equiv \det(\mathrm{I}_2 - 1_2\boldsymbol{s}^+_{t\cdot}) = 1 - \boldsymbol{s}^+_{t\cdot}1_2\ , \tag{47b}$$

(this can easily be verified by direct calculation) one can solve for $\boldsymbol{a}'_t$, according to

$$\boldsymbol{a}'_t = -\delta_t^{-1}\boldsymbol{s}^+_{t\cdot}\ \mathrm{diag}\left[[(\mu_3 w_{t3})/(\mu_1 w_{t1}), (\mu_3 w_{t3})/(\mu_2 w_{t2})]'\right]\ .$$

Now, $\mathrm{rank}[\mathbf{M\Psi}_3(\mathbf{W}' \otimes \mathrm{I}_3)\mathbf{A}]$ is equal to 2 if, and only if, there exists a set $\{\boldsymbol{a}'_1, \boldsymbol{a}'_2, \dots, \boldsymbol{a}'_T\}$ such that $\boldsymbol{a}'_1 = \boldsymbol{a}'_2 = \cdots = \boldsymbol{a}'_T$ or, equivalently, if

$$\delta_t^{-1}\cdot\boldsymbol{s}^+_{t\cdot} = \delta_1^{-1}\cdot\boldsymbol{s}^+_{1\cdot}\cdot(w_{t3}/w_{13})^{-1}\mathrm{diag}[(w_{t1}/w_{11}, w_{t2}/w_{12})'] \tag{48}$$

for $t = 2,\ldots,T$. By the definition (47b), δ_t is equal to the predicted cost share for the 3rd input in period t. Accordingly, the condition (48) can alternatively be formulated in terms of the following requirement on the individual cost shares

$$\frac{s^+_{ti}}{s^+_{t3}} = \frac{s^+_{1i}}{s^+_{13}} \frac{w_{13}}{w_{t3}} \frac{w_{ti}}{w_{1i}} \quad \text{for} \quad i = 1,2 \quad \text{and} \quad t = 2,\ldots,T,$$

which, in turn, is equivalent to the condition

$$\frac{\tilde{x}_{ti}}{\tilde{x}_{t3}} = \frac{\tilde{x}_{1i}}{\tilde{x}_{13}} \quad \text{for} \quad i = 1,2 \quad \text{and} \quad t = 2,\ldots,T.$$

Thus, the rows of Υ_{22} cannot be linearly dependent unless all the predicted factor proportions are constant over time. Such constancy requires that all relative prices are constant over time and that there is no technical change ; cf. (9). The probability of this event can safely be set to zero. Q.E.D.

Next, the rank of Υ_{11} will be considered. Two lemmas are first given which describe the structure of Υ_{11} .

LEMMA 6. *The matrix* $\left[\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\kappa}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \left[\frac{\partial g^\star}{\partial \boldsymbol{\kappa}}\right]' \boldsymbol{\Psi}_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}')\right]\mathbf{A}$ *involved in* Υ_{11} , *can be expressed in terms of* 1×2 *vectors such that the* (h,t)*th element is given by*

$$\boldsymbol{z}_{t\cdot}\left[g^\star_t \frac{\partial \mathbf{B}'}{\partial \kappa_h} + \frac{\partial g^\star_t}{\partial \kappa_h}\mathbf{B}'\right](\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}^+_{t\cdot}) , \quad h = 1,\ldots,8, \; t = 1,\ldots,T,$$

where $\boldsymbol{z}_{t\cdot}$ *denotes the* t*th row of the* $T \times 5$ *matrix* $\mathbf{Z}$, $g^\star_t$ *the* t*th element of the* $T \times 1$ *vector* $\boldsymbol{g}^\star$ and $\partial\mathbf{B}'/\partial\kappa_h$ *denotes the* 5×3 *matrix obtained by differentiating* $\mathbf{B}'$ *componentwise with respect to* κ_h .

PROOF. See Appendix C.

LEMMA 7. *The matrix* $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ *is given by*

$$(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})' =$$

$$= \begin{bmatrix} 1/\lambda & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/\lambda & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \ln\lambda/\lambda & 0 & -\ln\lambda & 1/\lambda & 0 & -1 & 0 & 0 & 0 & -1/\lambda & 0 & 1 & 0 & 0 & 0 \\ \ln\lambda/\lambda & \ln\lambda/\lambda & -2\ln\lambda & 0 & 1/\lambda & -1 & 1/\lambda & 0 & -1 & -1/\lambda & -1/\lambda & 2 & 0 & 0 & 0 \\ 0 & \ln\lambda/\lambda & -\ln\lambda & 0 & 0 & 0 & 0 & 1/\lambda & -1 & 0 & -1/\lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/\lambda & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/\lambda & -1 \\ q_1 & q_2 & q_3 & q_4 & q_5 & 0 & q_7 & q_8 & 0 & q_{10} & q_{11} & 0 & q_{13} & q_{14} & 0 \end{bmatrix}$$

where

$$q_1 = [(\gamma_{11} + \gamma_{12})\cdot(\lambda^{-1} - \ln\lambda) - \alpha_1\cdot\ln\lambda]/\lambda\,,$$

$$q_2 = [(\gamma_{12} + \gamma_{22})\cdot(\lambda^{-1} - \ln\lambda) - \alpha_2\cdot\ln\lambda]/\lambda\,,$$

$$q_3 = -(\gamma_{11} + 2\gamma_{12} + \gamma_{22})\,, \quad q_4 = -\gamma_{11}/\lambda\,, \quad q_5 = -\gamma_{12}/\lambda\,,$$

$$q_7 = -\gamma_{12}/\lambda\,, \quad q_8 = -\gamma_{22}/\lambda\,, \quad q_{10} = (\gamma_{11} + \gamma_{12})/\lambda\,,$$

$$q_{11} = (\gamma_{12} + \gamma_{22})/\lambda\,, \quad q_{13} = -\tau_1/\lambda\,, \quad q_{14} = -\tau_2/\lambda\,.$$

and has full row rank, i.e. 8, *except at points in the parameter space where* $\boldsymbol{\kappa} = (\alpha_1, \alpha_2, \gamma_{11}, \gamma_{12}, \gamma_{22}, \tau_1, \tau_2, \beta)'$ *can be written as* $(\alpha_1, -\alpha_1, \gamma_{11}, -\gamma_{11}, \gamma_{11}, \tau_1, -\tau_1, \beta)'$ or $(\alpha_1, \alpha_2, \gamma_{11}, -\gamma_{11}, \gamma_{11}, \tau_1, -\tau_1, 0)'$. *At these points the rank of* $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ *is equal to* 7.

PROOF. The explicit matrix expression is obtained by vectorization and differentiation of **B**, as given by (20). The rank property can be proved as follows. Start by considering the first seven rows of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$. Denote the corresponding matrix $(\partial\mathbf{B}^c/\partial\underline{\boldsymbol{\kappa}})'$. As this matrix is a function of λ only its full rank can be verified by assigning an arbitrary (positive) value to λ, forming the product $[(\partial\mathbf{B}^c/\partial\underline{\boldsymbol{\kappa}})'(\partial\mathbf{B}^c/\partial\underline{\boldsymbol{\kappa}})]$ and checking that its inverse exists. Accordingly the rank is at least seven. That it is exactly seven will be proved by showing that the last row can be obtained as a linear combination of the first seven rows. This is tedious but straightforward work. To see how it can be done, first consider the last three elements of the last row. It is clear that these must be obtained by means of the sixth and seventh rows. It can also be seen that if a linear combination of rows six and seven is to result in a vector with a zero last element, then it must be the case that $\tau_1 = -\tau_2$. Similarly, consider the fourth, fifth, and sixth elements of the last row (i.e q_4, q_5, and q_6). To obtain these elements rows three and four have to be weighted by $-\gamma_{11}$ and $-\gamma_{12}$, respectively, and, moreover, it must be the case that $\gamma_{12} = -\gamma_{11}$ in order to ascertain that $q_6 = 0$. The constraint $\gamma_{22} = -\gamma_{12}$ can be established in the same way by considering the elements q_7, q_8, and q_9. Thus, $\gamma_{22} = \gamma_{11}$. Further, note that, given these constraints, $q_1 = -\alpha_1 \cdot \ln\lambda/\lambda$, $q_2 = -\alpha_2 \cdot \ln\lambda/\lambda$, and $q_3 = 0$. Since rows three, four, and five are weighted by γ_{11}, $-\gamma_{11}$, and γ_{11}, respectively, rows one and two have to be weighted by $-\alpha_1 \cdot \ln\lambda$ and $-\alpha_2 \cdot \ln\lambda$ to generate q_1 and q_2. Finally, $q_3 = 0$ implies that the weights for rows one and two must sum to zero which implies that either $\alpha_2 = -\alpha_1$ or $\beta = 0$ (i.e $\lambda = 1$). Q.E.D.

Concerning Lemma 7, it is worth noting that $(\partial \mathbf{B}^c/\partial \boldsymbol{\kappa})'$ has full row rank for a static, allocatively inefficient Cobb-Douglas technology, i.e. if $\gamma_{11} = \gamma_{12} = \gamma_{22} = \tau_1 = \tau_2 = 0$, provided that either $\alpha_2 \neq -\alpha_1$ or $\beta \neq 0$. This indicates that full row rank on $(\partial \mathbf{B}^c/\partial \boldsymbol{\kappa})'$ cannot be a sufficient condition for identification, because any attempt to estimate a model of this type will fail, at least as long as technical efficiency is assumed. That can be inferred by mere inspection of the cost shares (11), remembering that only the two first shares will be considered in the estimation. The following theorem shows that the conditions under which $(\partial \mathbf{B}^c/\partial \boldsymbol{\kappa})'$ has full row rank is a special case of the conditions under which $\boldsymbol{\Upsilon}_{11}$ has full rank. As expected, the latter conditions imply that the Cobb-Douglas model just discussed is not identified. In fact, the theorem implies that identification will fail even if the technology modeled only has a partial Cobb-Douglas structure; this is further commented upon after the proof of the theorem.

THEOREM 3. *A sufficient condition for the* 8×8 *matrix* Υ_{11}, *given by* (41), *to have full rank is that the parameter vector* κ *does not belong to the subspace of the parameter space which satisfies*

$\kappa = (\alpha_1, \alpha_2, \gamma_{11}, \gamma_{12}, \gamma_{22}, \tau_1, \tau_2, \beta)' = (\alpha_1, \alpha_2, \gamma_{11}, -\gamma_{11}, \gamma_{11}, \tau_1, -\tau_1, \beta)'$.

If technical inefficiency is not modeled (i.e. assumed non-existent a priori) then this condition is also necessary.

PROOF. By (41), Lemma 4, and the results concerning ranks that were exploited in the proof of Lemma 4:

$$\begin{aligned}\operatorname{rank}(\Upsilon_{11}) &= \operatorname{rank}\Big\{\Big[\Big[\frac{\partial \mathbf{B}^c}{\partial \kappa}\Big]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \Big[\frac{\partial g^\star}{\partial \kappa}\Big]'\Psi_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}')\Big]\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P}) \\ &\quad \times [\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})]'\Big[\Big[\frac{\partial \mathbf{B}^c}{\partial \kappa}\Big]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \Big[\frac{\partial g^\star}{\partial \kappa}\Big]'\Psi_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}')\Big]'\Big\} \\ &= \operatorname{rank}\Big\{\Big[\Big[\frac{\partial \mathbf{B}^c}{\partial \kappa}\Big]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \Big[\frac{\partial g^\star}{\partial \kappa}\Big]'\Psi_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}')\Big]\mathbf{A}(\mathbf{D}^{-1} \otimes \mathbf{P})\Big\} \\ &= \operatorname{rank}\Big\{\Big[\Big[\frac{\partial \mathbf{B}^c}{\partial \kappa}\Big]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \Big[\frac{\partial g^\star}{\partial \kappa}\Big]'\Psi_T(\mathbf{I}_T \otimes \mathbf{Z}\mathbf{B}')\Big]\mathbf{A}\Big\}\end{aligned}$$

It will be shown that the rows of the matrix in the last equality cannot be linearly dependent.

From Lemma 6 it is clear that if the rows of Υ_{11} are linearly dependent then there exists a vector $\zeta \in \mathbb{R}^8$, $\zeta \neq \mathbf{0}_8$ such that

$$\boldsymbol{z}_{t\cdot}\Big[\sum_{h=1}^{8}\xi_h \cdot \Big[g_t^\star \frac{\partial \mathbf{B}'}{\partial \kappa_h} + \frac{\partial g_t^\star}{\partial \kappa_h}\mathbf{B}'\Big]\Big](\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+) = \mathbf{0}_2', \quad t = 1,\dots,T. \tag{49}$$

The first thing to examine is whether there are conditions under which the matrix within brackets may be equal to $\mathbf{0}_{5,3}$. Lemma 7 states the circumstances which make it possible to make $\Sigma_{h=1}^{8}\xi_h\cdot(\partial\mathbf{B}'/\partial\kappa_h)$ equal to $\mathbf{0}_{5,3}$. For the matrix within brackets to be singular these conditions should also imply that the sum $\Sigma_{h=1}^{8}\xi_h\cdot(\partial g_t^\star/\partial\kappa_h)$ equals zero. It can checked, however, that this is not the case.

Next, it must be checked if it is possible to satisfy (49) if the matrix in brackets is different from the zero matrix. Note that

$$(\boldsymbol{s}_{t\cdot}^{+}\ ,\ s_{t3}^{+})(\mathbf{J}' - \mathbf{1}_3\boldsymbol{s}_{t\cdot}^{+}) = (\boldsymbol{s}_{t\cdot}^{+}\ ,\ s_{t3}^{+})\begin{bmatrix}(\mathbf{I}_2 - \mathbf{1}_2\boldsymbol{s}_{t\cdot}^{+}) \\ -\ \boldsymbol{s}_{t\cdot}^{+}\end{bmatrix} = \mathbf{0}_2'\ ,$$

since $s_{t3}^{+} = 1 - \boldsymbol{s}_{t\cdot}^{+}\mathbf{1}_2$. Secondly, notice also that

$$(\boldsymbol{s}_{t\cdot}^{+}\ ,\ s_{t3}^{+}) = [(\boldsymbol{z}_{t\cdot}\mathbf{B}'g_t^\star + \boldsymbol{w}_{t\cdot}\mathbf{M})\mathbf{1}_3]^{-1}\cdot(\boldsymbol{z}_{t\cdot}\mathbf{B}'g_t^\star + \boldsymbol{w}_{t\cdot}\mathbf{M})\ ,$$

where the first factor on the RHS is a scalar, namely total expenditure divided by $f(y,t)$, and the second factor is the (row) vector of predicted expenditures divided by $f(y,t)$. Utilizing these two observations, one can reformulate (49) according to

$$\boldsymbol{z}_{t\cdot}\left[\sum_{h=1}^{8}\zeta_{th}\cdot\left[g_t^\star\cdot\frac{\partial\mathbf{B}'}{\partial\kappa_h} + \frac{\partial g_t^\star}{\partial\kappa_h}\mathbf{B}'\right]\right] = \boldsymbol{z}_{t\cdot}\mathbf{B}'g_t^\star + \boldsymbol{w}_{t\cdot}\mathbf{M}\ ,\quad t = 1,\ldots,T, \qquad (50)$$

where

$$\zeta_t = (\zeta_{t1},\ldots,\zeta_{t8})' \propto \xi = (\zeta_1,\ldots,\zeta_8)'$$

and " $\propto$ " denotes proportional to. Note that the implicit constant of proportionality may vary over time and, further, need not be equal to $[(\boldsymbol{z}_{t\cdot}\mathbf{B}'g_t^\star + \boldsymbol{w}_{t\cdot}\mathbf{M})\mathbf{1}_3]^{-1}$.

In general terms the LHS of (50) can be written $f(\boldsymbol{z}_{t\cdot}\ ;\ \boldsymbol{\kappa})$ whereas the RHS can be written as $g(\boldsymbol{z}_{t\cdot}\ ;\ \boldsymbol{\kappa}) + h(\boldsymbol{w}_{t\cdot}\ ;\ \boldsymbol{\nu})$. Since $\boldsymbol{w}_{t\cdot}$ cannot be written as a linear function of $\boldsymbol{z}_{t\cdot}$, and $\boldsymbol{\nu}$ and $\boldsymbol{\kappa}$ are not functionally related at all, a necessary condition for (50) to be satisfied is that the function $h(\boldsymbol{w}_{t\cdot}\,;\boldsymbol{\nu})$ is identically equal to zero. This can only happen if there is no technical inefficiency, in which case $\mathbf{M} = \mathbf{0}_{3,3}$ and $g_t^{\star}$ can be set equal to unity [cf. (22)] .[23] Put differently, if there is technical inefficiency then $\boldsymbol{\Upsilon}_{11}$ must have full rank. This is the reason for the last statement in the theorem; the mere modeling of technical inefficiency may suffice to give $\boldsymbol{\Upsilon}_{11}$ full rank.

It remains to consider the case with technical efficiency. To begin with, notice that since in this case the matrix within brackets in (49) is given by $\Sigma_{h=1}^{8}\zeta_h\cdot(\partial\mathbf{B}'/\partial\kappa_h)$ the conditions in Lemma 7 will be sufficient to satisfy (49). Accordingly, in the context of technical efficiency a necessary condition for the full rank of $\boldsymbol{\Upsilon}_{11}$ is that matrix $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ should have full row rank.

However, full row rank of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ is not sufficient to ascertain that $\boldsymbol{\Upsilon}_{11}$ has full rank; it has to be proved that, in addition, (50) can not be satisfied by any $\zeta_t \in \mathbb{R}^8$. Given technical efficiency, (50) can be expressed according to

$$\boldsymbol{z}_{t\cdot}\sum_{h=1}^{8}\zeta_{th}\cdot\frac{\partial\mathbf{B}'}{\partial\kappa_h} = \boldsymbol{z}_{t\cdot}\mathbf{B}' \;\Rightarrow\; \sum_{h=1}^{8}\zeta_h\cdot\frac{\partial\mathbf{B}'}{\partial\kappa_h} = \mathbf{B}' \;\Longleftrightarrow\; \frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\kappa}}\,\zeta = \mathbf{B}^c ,$$

where

$$\zeta_h \equiv T^{-1}\cdot\sum_{t=1}^{T}\zeta_{th}\,, \quad h = 1,\ldots,8.$$

The expression $\Sigma_{h=1}^{8}\zeta_h(\partial\mathbf{B}'/\partial\kappa_h) = \mathbf{B}'$ follows by multiplying both sides of the first equality by $\boldsymbol{z}_{t\cdot}'$, summing over t, dividing by T and, finally, multiplying both sides by $(\mathbf{Z}'\mathbf{Z})^{-1}$. What the condition says is that if the rows of $\boldsymbol{\Upsilon}_{11}$ are linearly dependent then it should be possible to express $\mathbf{B}^c$

---

[23] While it is conceivable that the required proportinality might be satisfied for some individual observation even in the presence of technical inefficiency, the probability of proportionality holding at *every* observation when $\mathbf{M} \neq \mathbf{0}$ can safely be set equal to zero.

as a linear combination of its first order derivatives with respect to $\boldsymbol{\kappa}$.

By means of the technique employed in the proof of Lemma 7, it will be shown that $\mathbf{B}^{c\prime}$ can be written as a linear combination of the rows of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ if, and only if, (i) the vector $\boldsymbol{\kappa} = (\alpha_1, \alpha_2, \gamma_{11}, \gamma_{12}, \gamma_{22}, \tau_1, \tau_2, \beta)'$ can be written $(\alpha_1, \alpha_2, \gamma_{11}, -\gamma_{11}, \gamma_{11}, \tau_1, -\tau_1, \beta)'$ and (ii) $(\alpha_1 + \alpha_2)\cdot\ln\lambda \neq 0$. Since the conditions on $\boldsymbol{\kappa}$ considered in Lemma 7 concern the cases when the product $(\alpha_1 + \alpha_2)\cdot\ln\lambda$ equals zero, the requirements (i) and (ii) plus Lemma 7 must imply the condition on $\boldsymbol{\kappa}$ that is stated in the theorem.

In calculating $\mathbf{B}^{c\prime}$ as a linear combination of the rows of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ note, first, that ζ_8 must be non-zero. This can be seen as follows. If ζ_8 is zero then it is easy to see that ζ_3, ζ_4, ζ_5, ζ_6 and ζ_7 must be γ_{11}, γ_{12}, γ_{22}, τ_1, and τ_2, respectively. Given these weights it also clear that to obtain b_{11} and b_{12} one must set ζ_1 and ζ_2 equal to α_1 and α_2, respectively. However, with the weights of rows 1, 2, 3, 4, and 5 thus given by α_1, α_2, γ_{11}, γ_{12}, and γ_{22} the third element in the linear combination vector will be

$$-\alpha_1 - \alpha_2 - (\gamma_{11} + 2\gamma_{12} + \gamma_{22})$$

while the third element of $\mathbf{B}^{c\prime}$ is

$$1 - \alpha_1 - \alpha_2 - (\gamma_{11} + 2\gamma_{12} + \gamma_{22})\,.$$

Clearly, the two can never be equal.

Consider, therefore, a linear combination in which $\zeta_8 \neq 0$. Inspection of the last three elements of $\mathbf{B}^{c\prime}$ immediately shows such combinations require that $\tau_2 = -\tau_1$ and that the weights of rows 6 and 7 must satisfy $\zeta_6 = (1 + \zeta_8)\cdot\tau_1$ and $\zeta_7 = -(1 + \zeta_8)\cdot\tau_1$, respectively, where ζ_8 is yet to be determined. Proceeding to the seventh, eighth and ninth elements of $\mathbf{B}^{c\prime}$, i.e. $b_{13} = \gamma_{12}/\lambda$, $b_{23} = \gamma_{22}/\lambda$, and $b_{33} = -(\gamma_{12} + \gamma_{22})$ one concludes

$$\gamma_{22} = -\gamma_{12}\,, \quad \zeta_4 = (1 + \zeta_8)\cdot\gamma_{12}\,, \quad \zeta_5 = -(1 + \zeta_8)\cdot\gamma_{12}\,.$$

Likewise, for the fourth, fifth and sixth elements of $\mathbf{B}^{c\prime}$, i.e. b_{21}, b_{22} and b_{32}:

$$\gamma_{12} = -\gamma_{11}\,, \quad \zeta_3 = (1+\zeta_8)\cdot\gamma_{11} \quad [\text{and } \zeta_4 = -(1+\zeta_8)\cdot\gamma_{11}\,]\,.$$

Summarizing, the analysis so far implies that

$$\gamma_{12} = -\gamma_{11}\,, \quad \gamma_{22} = -\gamma_{11}\,, \quad \tau_2 = -\tau_1 \tag{51a}$$

$$\begin{aligned} &\zeta_3 = (1+\zeta_8)\cdot\gamma_{11}\,, \quad \zeta_4 = -\zeta_3\,, \quad \zeta_5 = \zeta_3\,, \\ &\zeta_6 = (1+\zeta_8)\cdot\tau_1\,, \quad \zeta_7 = -\zeta_6\,. \end{aligned} \tag{51b}$$

Three weights remain to be determined, namely, ζ_1, ζ_2 and ζ_3. The information that $\gamma_{12} = -\gamma_{11}$ and $\gamma_{22} = \gamma_{11}$ makes it possible to formulate the following three equation system by means of the three first columns of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ and the three first elements of $\mathbf{B}^{c\prime}$:

$$\begin{bmatrix} 1 & 0 & -\alpha_1\cdot\ln\lambda \\ 0 & 1 & -\alpha_2\cdot\ln\lambda \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_8 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ 1-\alpha_1-\alpha_2 \end{bmatrix}.$$

The determinant of the system is given by $D = -(\alpha_1+\alpha_2)\cdot\ln\lambda$; hence the requirement (ii) above that $(\alpha_1+\alpha_2)\cdot\ln\lambda \neq 0$. Imposing this condition and solving the system one obtains

$$\begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_8 \end{bmatrix} = -\frac{1}{(\alpha_1+\alpha_2)\cdot\ln\lambda} \begin{bmatrix} \alpha_1(1-\alpha_1-\alpha_2)\cdot\ln\lambda \\ \alpha_2(1-\alpha_1-\alpha_2)\cdot\ln\lambda \\ 1 \end{bmatrix}.$$

It can easily be checked that if this result and (51) are applied to the rows of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ then $\mathbf{B}^{c\prime}$ results. Accordingly, $\boldsymbol{\Upsilon}_{11}$ cannot have full rank under the specified restrictions [i.e. (51a) plus the condition $(\alpha_1+\alpha_2)\cdot\ln\lambda \neq 0$]. Since these restrictions were necessary to obtain *any* solution and since the solution actually obtained is unique there can be no other linear combination of the rows of $(\partial\mathbf{B}^c/\partial\boldsymbol{\kappa})'$ that yields $\mathbf{B}^{c\prime}$. This completes the proof. Q.E.D.

To provide some intuition for the conditions under which $\boldsymbol{\Upsilon}_{11}$ is singular note that

$$\gamma_{12} = -\gamma_{11} \implies \gamma_{13} = -(\gamma_{11} + \gamma_{12}) = 0 \implies \sigma_{13} = 1 ,$$

and

$$\gamma_{22} = -\gamma_{12} \implies \gamma_{23} = -(\gamma_{12} + \gamma_{22}) = 0 \implies \sigma_{23} = 1 ,$$

where σ_{13} and σ_{23} denote the Allen partial elasticities of substitution between input 1 and input 3 and between input 2 and input 3, respectively. Accordingly, the information matrix may be singular if the model has partial Cobb-Douglas structure in that input 3 has a Cobb-Douglas relation to the two other inputs. This requires, however, that $\tau_2 = -\tau_1$ [and, hence, that $\tau_3 = -(\tau_1 + \tau_2) = 0$]. Still, this means that the condition in Theorem 3 is not merely a mathematical subtlety. For an example of an empirical study based on the translog cost function in which a partial Cobb-Douglas structures could not be rejected, see Berndt and Wood (1975).

It remains to prove that $(\boldsymbol{\Upsilon}_{11} - \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}'_{12})$ has full rank. This is done in the proof of the following theorem.

THEOREM 4. *The* 8×8 *matrix* $(\boldsymbol{\Upsilon}_{11} - \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}'_{12})$, *where* $\boldsymbol{\Upsilon}_{11}$, $\boldsymbol{\Upsilon}_{12}$, *and* $\boldsymbol{\Upsilon}_{22}$ *are given by* (41), (42), and (43), *respectively, has full rank if* $\boldsymbol{\Upsilon}_{11}$ *has full rank, i.e. under the conditions stated in Theorem* 3.

PROOF. For simplicity, define

$$\mathbf{C} \equiv \left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\kappa}}\right]'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3) + \left[\frac{\partial \boldsymbol{g}^{\star}}{\partial \boldsymbol{\kappa}}\right]'\boldsymbol{\Psi}_{\mathrm{T}}(\mathbf{I}_{\mathrm{T}} \otimes \mathbf{Z}\mathbf{B}') \tag{52}$$

and

$$\mathbf{H} \equiv (\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'(\mathbf{W} \otimes \mathbf{I}_3)\boldsymbol{\Psi}'_3\mathbf{M} . \tag{53}$$

By means of these definitions and Lemma 4, Υ_{11}, Υ_{12}, and Υ_{22} can be written

$$\Upsilon_{11} = \mathbf{CA}(\mathbf{D}^{-1} \otimes \mathbf{P})(\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'\mathbf{C}' \tag{54}$$

$$\Upsilon_{12} = \mathbf{CA}(\mathbf{D}^{-1} \otimes \mathbf{P})\mathbf{H}$$

$$\Upsilon_{22} = \mathbf{H}'\mathbf{H}\,.$$

Consequently,

$$\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}' = \mathbf{CA}(\mathbf{D}^{-1} \otimes \mathbf{P})[\mathbf{I}_{\mathrm{Tq}} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'](\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'\mathbf{C}'\,.$$

Note that $[\mathbf{I}_{\mathrm{Tq}} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}']$ is symmetric and idempotent. Thus, since $\mathrm{rank}(\mathbf{XX}') = \mathrm{rank}(\mathbf{X}')$,

$$\begin{aligned}\mathrm{rank}(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}') &= \mathrm{rank}\{[\mathbf{I}_{\mathrm{T}\cdot 2} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'](\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'\mathbf{C}'\} \\ &\leq \min\,\{\mathrm{rank}[\mathbf{I}_{\mathrm{T}\cdot 2} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'],\, \mathrm{rank}[(\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'\mathbf{C}']\}\end{aligned}$$

By Theorem 3 and (54), $\mathrm{rank}[(\mathbf{D}^{-1} \otimes \mathbf{P}')\mathbf{A}'\mathbf{C}'] = 8$. The rank of the matrix $[\mathbf{I}_{\mathrm{T}\cdot 2} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}']$ can be determined as follows. Notice, first, that since $\mathbf{H}$ has full column rank (this follows from the fact that Υ_{22} has full rank) the Moore - Penrose inverse of $\mathbf{H}$ is given by $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$. Thus, application of Theorem 8 in Magnus and Neudecker (1988, p. 35) yields

$$\mathrm{rank}\{\mathbf{I}_{\mathrm{T}\cdot 2} - \mathbf{H}[(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}']\} = \mathrm{rank}(\mathbf{I}_{\mathrm{T}\cdot 2}) - \mathrm{rank}(\mathbf{H}) = T\cdot 2 - 3\,. \tag{55}$$

Assuming T to be at least equal to the number of parameters to be estimated $T\cdot 2 - 3$ will always be strictly greater than 8. It might thus be conjectured $(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}')$ has full rank. That this conjecture is correct will be proved by showing that $\mathrm{rank}(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}')$ cannot be less than 8.

Accordingly, assume that the rank of $(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}')$ is less than 8.

This means that there exists an 8×1 vector ξ such that

$$\mathrm{I}_{\mathrm{T}\cdot 2} - \mathrm{H}(\mathrm{H'H})^{-1}\mathrm{H'}](\mathrm{D}^{-1} \otimes \mathrm{P'})\mathrm{A'C'}\xi = \mathbf{0}_{\mathrm{T}\cdot 2}\,; \quad \xi \in \mathbb{R}^{8}, \quad \xi \neq \mathbf{0}_{8}\,. \tag{56}$$

Remember that $(\mathrm{D}^{-1} \otimes \mathrm{P'})\mathrm{A'C'}$ has full column rank and note that the null space of $[\mathrm{I}_{\mathrm{T}\cdot 2} - \mathrm{H}(\mathrm{H'H})^{-1}\mathrm{H'}]$ is spanned by H. The latter fact is easily established by means of the equality $[\mathrm{I}_{\mathrm{T}\cdot 2} - \mathrm{H}(\mathrm{H'H})^{-1}\mathrm{H'}]\mathrm{H} = \mathbf{0}_{\mathrm{T}\cdot 2,3}$ and (55).

Thus, if true, (56) implies that there exists an 3×1 vector η such that

$$(\mathrm{D}^{-1} \otimes \mathrm{P'})\mathrm{A'C'}\xi \;=\; -\mathrm{H}\eta; \qquad \eta \in \mathbb{R}^{3}, \quad \eta \neq \mathbf{0}_{3}\,.$$

Premultiplying by $(\mathrm{D} \otimes \mathrm{P'}^{-1})$, using (52) and (53) and rearranging one obtains

$$\left\{\xi'\left[\left[\frac{\partial \mathrm{B}^{c}}{\partial \boldsymbol{\kappa}}\right]'(\mathrm{Z'G} \otimes \mathrm{I}_{3}) + \left[\frac{\partial g^{\star}}{\partial \boldsymbol{\kappa}}\right]'\Psi_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}} \otimes \mathrm{ZB'})\right] + \eta'\mathrm{M}\Psi_{3}(\mathrm{W'} \otimes \mathrm{I}_{3})\right\}\mathrm{A} = \mathbf{0}'_{\mathrm{T}\cdot 2}\,. \tag{57}$$

Note that (57) can be seen as a generalization of the conditions examined in Theorem 2 and Theorem 3; in Theorem 2 it is shown that (57) cannot be true if $\xi = \mathbf{0}_{8}$ and $\eta \neq \mathbf{0}_{3}$ while in Theorem 3 it is demonstrated that (57) is not fulfilled for $\xi \neq \mathbf{0}_{8}$ and $\eta = \mathbf{0}_{3}$.

Reformulating (57) in analogy with the condition (49) used in the proof of Theorem 3 one obtains

$$\left[\boldsymbol{z}_{\mathrm{t}\cdot}\left[\sum_{\mathrm{h}=1}^{8}\xi_{\mathrm{h}}\cdot g^{\star}_{\mathrm{t}}\cdot\frac{\partial \mathrm{B'}}{\partial \kappa_{\mathrm{h}}} + \frac{\partial g^{\star}_{\mathrm{t}}}{\partial \kappa_{\mathrm{h}}}\mathrm{B'}\right] + \eta'\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}_{\mathrm{t}\cdot}')\right](\mathrm{J'} - \mathbf{1}_{3}\boldsymbol{s}^{+}_{\mathrm{t}\cdot}) = \mathbf{0}'_{2}\,.$$

The matrix $\mathrm{diag}(\boldsymbol{\mu} \odot \boldsymbol{w}_{\mathrm{t}\cdot}')$ is obviously non-singular.[24] Further, in the proof of Theorem 3 it is was shown that there are no conditions under which the matrix within the large parentheses may be equal to the 5×3 zero matrix. Accordingly, it is impossible to satisfy the condition by making the expression

[24] Of course, it is assumed here that technical inefficiency is modeled, rather than taken to be non-existent a priori. Otherwise, the present theorem would be irrelevant.

within brackets equal to a sum of two zero vectors. Neither is it possible that $\eta'\text{diag}(\mu \odot w_{t.}') = -z_{t.}\left[\cdot\right]$ for all $t = 1,\ldots,T$, because $w_{t.}$ not a linear function of $z_{t.}$ and the parameters in κ and μ are not functionally related. Hence, the condition cannot satisfied by making the sum of the vectors within the brackets equal to a zero vector.

Accordingly, it remains to examine if the equality can be fulfilled when the two vectors within the brackets are non-zero and (non-trivially) distinct. To this end, (57) is reformulated in analogy with the condition (50) used in the proof of Theorem 3, yielding the following condition which is satisfied if $(\Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{12}')$ does not have full rank

$$z_{t.}\left[\sum_{h=1}^{8}\zeta_{th}\cdot g_t^{\star}\cdot\frac{\partial \mathbf{B}'}{\partial \kappa_h} + \frac{\partial g_t^{\star}}{\partial \kappa_h}\mathbf{B}'\right] + \rho_t'\text{diag}(\mu \odot w_{t.}') = z_{t.}\mathbf{B}'g_t^{\star} + w_{t.}\mathbf{M} \quad (58)$$

for all $t = 1,\ldots,T$ where

$$\zeta_t = (\zeta_{t1},\ldots,\zeta_{t8})' \propto \xi = (\xi_1,\ldots,\xi_8)' ,$$

$$\rho_t = (\rho_{t1}, \rho_{t2}, \rho_{t3})' \propto \eta = (\eta_1, \eta_2, \eta_3)' .$$

The first thing to notice is that $\rho_t'\text{diag}(\mu \odot w_{t.}')$ can always be made to be equal to $w_{t.}\mathbf{M}$ simply by letting $\rho_t' = (1, 1, 1)$.

Next consider the possibilities to choose ζ_t such that

$$z_{t.}\left[\sum_{h=1}^{8}\zeta_{th}\cdot\left[g_t^{\star}\cdot\frac{\partial \mathbf{B}'}{\partial \kappa_h} + \frac{\partial g_t^{\star}}{\partial \kappa_h}\mathbf{B}'\right]\right] = z_{t.}\mathbf{B}'g_t^{\star} .$$

To simplify this equation, premultiply both sides by $(g_t^{\star})^{-1}\cdot z_{t.}$, sum over t, divide by T, and, finally, premultiply both sides by $(\mathbf{Z}'\mathbf{Z})^{-1}$. This yields the following condition:

$$\sum_{h=1}^{8}\zeta_h\cdot\frac{\partial \mathbf{B}'}{\partial \kappa_h} + \mathbf{B}'\sum_{h=1}^{8}\zeta_h\cdot\left[\frac{1}{T}\cdot\sum_{t=1}^{T}\frac{\partial g_t^{\star}/\partial \kappa_h}{g_t^{\star}}\right] = \mathbf{B}' , \quad (59)$$

where

$$\zeta_h \equiv T^{-1} \cdot \sum_{t=1}^{T} \zeta_{th} , \qquad h = 1,\dots,8 .$$

Some reflection shows that (59) is satisfied under precisely the conditions which make the matrix Υ_{11} singular. That is to say, (59) holds whenever there exists a vector $\tilde{\zeta}$ such that

$$\sum_{h=1}^{8} \tilde{\zeta}_h \cdot \frac{\partial \mathbf{B}'}{\partial \kappa_h} = \mathbf{B}' \tag{60}$$

or a vector $\hat{\zeta}$ such that

$$\sum_{h=1}^{8} \hat{\zeta}_h \cdot \frac{\partial \mathbf{B}'}{\partial \kappa_h} = \mathbf{0}_{5,3} \tag{61}$$

To see this, note that if (60) holds then (59) can be satisfied by means of the following choice of the ζ_h's :

$$\zeta_h = \tilde{\zeta}_h \cdot (1 + \phi)^{-1} ; \quad \phi \equiv \sum_{h=1}^{8} \tilde{\zeta}_h \cdot \left[\frac{1}{T} \cdot \sum_{t=1}^{T} \frac{\partial g_t^{\star} / \partial \kappa_h}{g_t^{\star}} \right] , \quad h = 1,\dots,8.$$

On the other hand, if (61) is true then the ζ_h's can be chosen according to

$$\zeta_h = \hat{\zeta}_h \cdot \psi^{-1} ; \quad \psi \equiv \sum_{h=1}^{8} \hat{\zeta}_h \cdot \left[\frac{1}{T} \cdot \sum_{t=1}^{T} \frac{\partial g_t^{\star} / \partial \kappa_h}{g_t^{\star}} \right] , \quad h = 1,\dots,8.$$

Since it is clear that (59) can be satisfied only under either of the conditions (60) or (61) the Theorem follows. Q.E.D.

7. Concluding comments

The above analysis has shown that in the context of a static, three input translog technology, subject to non-neutral technical change, it is possible to identify both allocative and technical inefficiency parametrically. It was also demonstrated, however, that there exists a set of parameter constellations for which the model will not be identified. This conclusion was arrived at by extending the general results of Berndt et al. (1974) for non-linear multivariate regression models to the system of translog input cost share system and, subsequently, applying Rothenberg's (1971) identification criterion.

It is interesting to note that the existence of subspace of the parameter space where the model is not identified is due to the parametric specification of allocative inefficiency. As mentioned earlier this specification has been extensively used in the literature. Apparently, it has not been noted before that the class of models for which identification fails comprises, e.g., all static models with a *partial* Cobb-Douglas structure. Of course, this class is much larger than the class of (full) Cobb-Douglas models for which the lack of identification is trivial to establish. The formulation of technical inefficiency which has been used here, and which is new, does not induce any identification problems (beyond the general problem of increasing the number of parameters to be estimated by means of a given data set.)

Concerning the possibilities to generalize the results obtained in this paper it should be said that the restriction (7c) which reduces the number of parameters associated with allocative inefficiency from two to one, greatly simplifies the calculations in Section 6. To substitute other restrictions which have same effect for (7a) – e.g. that two of the three parameters be equal to unity a priori – poses no problem. As soon as only (7a) and (7b) are imposed it becomes much harder to derive the conditions under which the model is identified, at least by means of the techniques used here. However, there is no reason to believe that the basic result, i.e. that the model is identified for a large set of parameter constellations, should be overthrown. The same can be said about extensions involving a larger number of inputs.

Appendix A: Matrix operators and operations

This appendix contains a listing of some matrix results used in the paper. More extensive treatments can be found in Pollock (1979), Magnus (1988) and Magnus and Neudecker (1979, 1988).

A $m \times n$ matrix $\mathbf{X}$ may be regarded as an array of column vectors, $\boldsymbol{x}_{\cdot j}$, $j = 1,....,n$, or as an array of row vectors $\boldsymbol{x}_{i \cdot}$, $i = 1,....,m$. Thus, $\mathbf{X}$ may be vectorized either by stacking its columns on top of each other

$$\mathbf{X}^{c} = \begin{bmatrix} \boldsymbol{x}_{\cdot 1} \\ \vdots \\ \boldsymbol{x}_{\cdot n} \end{bmatrix} = (\boldsymbol{x}'_{\cdot 1}, ..., \boldsymbol{x}'_{\cdot n})', \tag{A.1}$$

or by putting its rows after one another

$$\mathbf{X}^{r} = (\boldsymbol{x}_{1 \cdot}, ..., \boldsymbol{x}_{m \cdot}) . \tag{A.2}$$ [25]

For a column vector, $\boldsymbol{y}$ say, (A.1) and (A.2) specialize into

$$\boldsymbol{y}^{c} = \boldsymbol{y}'^{c} = \boldsymbol{y} \tag{A.1'}$$

and

$$\boldsymbol{y}^{r} = \boldsymbol{y}'^{r} = \boldsymbol{y}' , \tag{A.2'}$$

respectively. The operators c and r are related according to

$$\mathbf{X}^{r} = \mathbf{X}'^{c'} \tag{A.3a}$$

$$\mathbf{X}'^{r} = \mathbf{X}^{c'} \tag{A.3b}$$

$$\mathbf{X}'^{r'} = \mathbf{X}^{c} \tag{A.3c}$$

[25] The use of the indices c and r is due to Pollock (1979). A more common notation for the column vector corresponding to $\mathbf{X}$ is vec($\mathbf{X}$) ; see, e.g., Magnus (op. cit.) and Magnus and Neudecker (op. cit.). If the vec notation is adopted there is no counterpart to the superindex r, however. For this reason Pollock's notation is often more convenient.

The commutation matrix $\mathbf{K}_{mn}$ is the $mn \times mn$ matrix which transforms $\mathbf{X}^c$ and $\mathbf{X}^r$ according to

$$\mathbf{K}_{mn}\mathbf{X}^c = \mathbf{X}'^c , \tag{A.4}$$

$$\mathbf{X}^r\mathbf{K}_{mn} = \mathbf{X}'^r . \text{ [26]} \tag{A.5}$$

The last result can be derived by means of (A.3a), (A.3b) and the following fact

$$\mathbf{K}'_{mn} = \mathbf{K}_{nm} = \mathbf{K}^{-1}_{mn} . \tag{A.6}$$

Let $\mathbf{A} = (a_{ij})$ be $r \times m$ and $\mathbf{B} = (b_{ij})$ $s \times n$. The Kronecker product is then the $rs \times mn$ matrix

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B}) \tag{A.7}$$

Some properties of Kronecker products are

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}) , \tag{A.8}$$

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{C}) , \tag{A.9}$$

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}' , \tag{A.10}$$

$$\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A}) \cdot \text{rank}(\mathbf{B}) , \tag{A.11}$$

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} . \tag{A.12}$$

The relationships between the operators c and r and the Kronecker product are given by

$$(\mathbf{AXB}')^c = (\mathbf{B} \otimes \mathbf{A})\mathbf{X}^c , \tag{A.13}$$

$$(\mathbf{A}'\mathbf{XB})^r = \mathbf{X}^r(\mathbf{A} \otimes \mathbf{B}) , \tag{A.14}$$

[26] The name commutation matrix and the denotation $\mathbf{K}_{mn}$ were introduced by Magnus and Neudecker (1979), where an explicit definition of $\mathbf{K}_{mn}$ can be found. The same matrix was defined in Pollock (1979) who called it the tensor commutator and denoted it by an encircled T. The former notation is used here because it is simpler and shows the dimension of the matrix.

while the following equality shows how the Kronecker product can be reversed by means of the commutation matrix

$$\mathbf{K}_{rs}(\mathbf{B} \otimes \mathbf{A}) = (\mathbf{A} \otimes \mathbf{B})\mathbf{K}_{mn} . \tag{A.15}$$

Let $\mathbf{A}$ and $\mathbf{C}$ be $m \times n$. The Hadamard product is then the $m \times n$ matrix

$$\mathbf{A} \odot \mathbf{C} = (a_{ij}c_{ij}) . \tag{A.16}$$

Finally, some definitions and results relating to square matrices. Denote the $n \times n$ diagonal matrix corresponding to the $n \times 1$ vector $\boldsymbol{y}$ by diag($\boldsymbol{y}$), i.e.

$$\operatorname{diag}(\boldsymbol{y}) = (\delta_{ij}y_i) , \qquad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} . \tag{A.17}$$

Further, denote by superindex d the operator which picks the diagonal elements of the $n \times n$ matrix $\mathbf{A}$ and arranges them into an $n \times 1$ vector :

$$\mathbf{A}^{\mathrm{d}} = (a_{11}, a_{22} ,...., a_{nn})' \qquad [\text{for } \mathbf{A} \;\; n \times n].\,^{27} \tag{A.18}$$

Thus, in particular,

$$[\operatorname{diag}(\boldsymbol{y})]^{\mathrm{d}} = \boldsymbol{y} . \tag{A.19}$$

The $n \times n^2$ basis matrix for diagonality $\boldsymbol{\Psi}_n$ is (implicitly) defined by

$$\boldsymbol{\Psi}'_n\boldsymbol{y} = [\operatorname{diag}(\boldsymbol{y})]^{\mathrm{c}} .\,^{28} \tag{A.20}$$

It can be shown that [cf. Magnus (1988, p. 110)]

$$\boldsymbol{\Psi}_n\mathbf{A}^{\mathrm{c}} = \mathbf{A}^{\mathrm{d}} \qquad [\text{for } \mathbf{A} \;\; n \times n] . \tag{A.21}$$

27 This notation is of my own making. Magnus (1988, p. 108) denotes this operator by w(A). However, given the use of the superindices c and r, it seems natural to use a superindex here, too. [The operation (A.18) is not considered in Pollock (1979) so there is no notation to be gotten from that source.]

28 An explicit definition of $\boldsymbol{\Psi}_n$ can be found in Magnus (1988, p. 109).

Appendix B: Results from matrix differential calculus

Following Pollock (1979) and Magnus and Neudecker (1988) matrix differentiation is treated within the framework of vector differentiation. According to the basic definition, the partial derivative of the $m \times 1$ vector $\boldsymbol{y}(\boldsymbol{x})$ with respect to the $n \times 1$ vector $\boldsymbol{x}$ is the $m \times n$ matrix

$$\partial \boldsymbol{y} / \partial \boldsymbol{x} = (\partial y_i / \partial x_j) \ . \tag{B.1}$$

Special cases of (B.1) are the derivative of a scalar, y, with respect to an $n \times 1$ vector $\boldsymbol{x}$, and the derivative of an $m \times 1$ vector $\boldsymbol{y}$ with respect to a scalar, x. These are given by the row vector

$$\partial y / \partial \boldsymbol{x} = (\partial y / \partial x_1 \ , \ .... \ , \ \partial y / \partial x_n) \tag{B.1'}$$

and the column vector

$$\partial \boldsymbol{y} / \partial x = (\partial y_1 / \partial x \ , \ .... \ , \ \partial y_m / \partial x)' \ , \tag{B.1''}$$

respectively. Regarding matrices, let $\mathbf{Y} = \mathbf{Y}(\mathbf{X})$ be $r \times s$ and $\mathbf{X}$ $m \times n$. Expressing the mapping in vector form, i.e. $\mathbf{Y}^c = \mathbf{Y}^c(\mathbf{X}^c)$, the derivative can be written

$$\partial \mathbf{Y}^c / \partial \mathbf{X}^c = (\partial \boldsymbol{y}_{.i} / \partial \boldsymbol{x}_{.j}) \ . \tag{B.2}$$

Notice that in this case the typical element is an $r \times m$ matrix.

Using (A.12) and (B.2)

$$\partial (\mathbf{AXB}')^c / \partial \mathbf{X}^c = (\mathbf{B} \otimes \mathbf{A})(\partial \mathbf{X}^c / \partial \mathbf{X}^c) = \mathbf{B} \otimes \mathbf{A} \ . \tag{B.3}$$

Further, by (A.4)

$$\partial \mathbf{X}'^c / \partial \mathbf{X}^c = \mathbf{K}_{mn} \ . \tag{B.4}$$

Let $\mathbf{Y} = \mathbf{Y}(\mathbf{U})$ and $\mathbf{U} = \mathbf{U}(\mathbf{X})$. Then, by the chain rule,

$$\partial\mathbf{Y}^c/\partial\mathbf{X}^c = (\partial\mathbf{Y}^c/\partial\mathbf{U}^c)(\partial\mathbf{U}^c/\partial\mathbf{X}^c) \,. \tag{B.5}$$

There is also an matrix analogue to the product rule. If $\mathbf{Y} = \mathbf{UVW}$ is an $r \times s$ matrix wherein $\mathbf{U} = \mathbf{U}(\mathbf{X}^c)$, $\mathbf{V} = \mathbf{V}(\mathbf{X}^c)$, and $\mathbf{W} = \mathbf{W}(\mathbf{X}^c)$ then

$$\begin{aligned}\partial(\mathbf{UVW})^c/\partial\mathbf{X}^c = {} & [(\mathbf{VW})' \otimes \mathbf{I}_r](\partial\mathbf{U}^c/\partial\mathbf{X}^c) + (\mathbf{W}' \otimes \mathbf{U})(\partial\mathbf{V}^c/\partial\mathbf{X}^c) \\ & + (\mathbf{I}_s \otimes \mathbf{UV})(\partial\mathbf{V}^c/\partial\mathbf{X}^c) \,,\end{aligned} \tag{B.6}$$

where $\mathbf{I}_r$ and $\mathbf{I}_s$ denote the identity matrices of order r and s, respectively. Of course, (B.5) and (B.6) can easily be specialized to the case when the derivative is taken with respect to an ordinary vector, z say, rather than $\mathbf{X}^c$; z is then simply substituted for $\mathbf{X}^c$. This is the case which will be encountered in the text.

By means of the above results one can derive the following two rules :

$$\partial\mathbf{U}^{-1c}/\partial\mathbf{U}^c = -\,(\mathbf{U}^{-1\prime} \otimes \mathbf{U}^{-1}) \tag{B.7}$$

and

$$\partial\ln|\mathbf{U}|/\partial\mathbf{U}^c = \mathbf{U}^{-1r} \,, \tag{B.8}$$

where $\ln|\mathbf{U}|$ denotes the natural logarithm of the determinant of $\mathbf{U}$.

Finally, two results concerning the derivatives of matrix traces, which can be found in Pollock (1979, p. 82)

$$\frac{\partial\mathrm{tr}(\mathbf{A}'\mathbf{A}\mathbf{X})^c}{\partial\mathbf{X}^c} = (\mathbf{A}'\mathbf{A})^r \,, \tag{B.9}$$

$$\frac{\partial\mathrm{tr}(\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{B}')^c}{\partial\mathbf{X}^c} = \mathbf{X}^{c\prime}(\mathbf{B} \otimes \mathbf{A}) + \mathbf{X}^{c\prime}(\mathbf{B}' \otimes \mathbf{A}') \,. \tag{B.10}$$

Appendix C: Proofs of lemmas

Proof of Lemma 1

Application of (B.6), (B.7) and (A.9) yields

$$\frac{\partial(\mathbf{S}^{+\prime})^c}{\partial\boldsymbol{\theta}} = (\mathbf{D}^{-1}\mathbf{GZ}\otimes\mathbf{J})\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\theta}} + (\mathbf{D}^{-1}\otimes\mathbf{JBZ}')\frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\theta}}$$
$$+ (\mathbf{D}^{-1}\mathbf{W}\otimes\mathbf{J})\frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\theta}} - (\mathbf{D}^{-1}\otimes\mathbf{S}^{+\prime})\frac{\partial\mathbf{D}^c}{\partial\boldsymbol{\theta}}\,.$$

The partial derivative $\partial\mathbf{D}^c/\partial\boldsymbol{\theta}$ can be expressed in terms of the partials $\partial\mathbf{B}^c/\partial\boldsymbol{\theta}$, $\partial\mathbf{G}^c/\partial\boldsymbol{\theta}$, and $\partial\mathbf{M}^c/\partial\boldsymbol{\theta}$. By (28), the definition (A.20) of the basis matrix for diagonality, and (A.1')

$$\mathbf{D}^c = \boldsymbol{\Psi}_T'(\mathbf{1}_3'\mathbf{BZ}'\mathbf{G} + \mathbf{1}_3'\mathbf{MW}')' = \boldsymbol{\Psi}_T'(\mathbf{1}_3'\mathbf{BZ}'\mathbf{G} + \mathbf{1}_3'\mathbf{MW}')^c\,.$$

Hence, by renewed application of the chain and product rules

$$\frac{\partial\mathbf{D}^c}{\partial\boldsymbol{\theta}} = \boldsymbol{\Psi}_T'\left[(\mathbf{GZ}\otimes\mathbf{1}_3')\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\theta}} + (\mathbf{I}_T\otimes\mathbf{1}_3'\mathbf{BZ}')\frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\theta}} + (\mathbf{W}\otimes\mathbf{1}_3')\frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\theta}}\right].$$

Inserting this result in the above expression for $\partial(\mathbf{S}^{+\prime})^c/\partial\boldsymbol{\theta}$ and rearranging, using (A.8) and (A.10), one obtains the first part of the lemma.

The second part follows from the definitions of **B**, **M**, and **G** [cf. Example 1, (21), and (28)] and from the definition (A.20). These imply

$$\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\theta}} = \left[\frac{\partial\mathbf{B}^c}{\partial\boldsymbol{\kappa}} \vdots \mathbf{0}_{5\cdot3,3}\right],$$
$$\frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\theta}} = \left[\mathbf{0}_{3^2,8} \vdots \frac{\partial\mathbf{M}^c}{\partial\boldsymbol{\nu}}\right] = \left[\mathbf{0}_{3^2,8} \vdots \boldsymbol{\Psi}_3'\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\nu}}\right] = \left[\mathbf{0}_{3^2,8} \vdots \boldsymbol{\Psi}_3'\mathbf{M}\right],$$
$$\frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\theta}} = \left[\frac{\partial\mathbf{G}^c}{\partial\boldsymbol{\kappa}} \vdots \mathbf{0}_{T^2,3}\right] = \left[\boldsymbol{\Psi}_T'\frac{\partial\boldsymbol{g}^\star}{\partial\boldsymbol{\kappa}} \vdots \mathbf{0}_{T^2,3}\right].$$

Q.E.D.

Proof of Lemma 2

By (B.5), (B.6), and (B.7) :

$$-\frac{\partial[\partial L(\boldsymbol{\theta},\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^c]'}{\partial\boldsymbol{\Omega}^c} = \tfrac{1}{2}\cdot[\boldsymbol{\Omega}^{-1}(\mathbf{R}-\mathbf{S}^+)'(\mathbf{R}-\mathbf{S}^+)\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}]$$
$$+\tfrac{1}{2}\cdot[\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}(\mathbf{R}-\mathbf{S}^+)'(\mathbf{R}-\mathbf{S}^+)\boldsymbol{\Omega}^{-1}] - \tfrac{\mathrm{T}}{2}\cdot(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\,.$$

Noting that, when evaluated at $(\boldsymbol{\theta}_0,\boldsymbol{\Omega}_0)$, the expectation of $(\mathbf{R}-\mathbf{S}^+)'(\mathbf{R}-\mathbf{S}^+)$ equals $\mathrm{T}\cdot\boldsymbol{\Omega}$ one obtains the desired expression. The rank of $\mathcal{I}(\boldsymbol{\Omega}_0)$ follows immediately from (A.11). Q.E.D.

Proof of Lemma 3

To prove (i), first consider the product $\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$. As shown by Magnus (1988, p. 109) the matrix $\boldsymbol{\Psi}_{\mathrm{T}}$ can be written

$$\boldsymbol{\Psi}_{\mathrm{T}} = [\mathrm{E}_{11}(\mathrm{T})\,,\,\mathrm{E}_{22}(\mathrm{T})\,,\,....\,,\,\mathrm{E}_{\mathrm{TT}}(\mathrm{T})]$$

where the submatrices $\mathrm{E}_{tt}(\mathrm{T})$, $t = 1,...,T$, are $T\times T$ with the (t,t)'th element equal to unity and all other elements equal to zero. This implies that the $T\times 2$ matrix $\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$ has the following structure:

$$\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+) = \begin{bmatrix} \boldsymbol{s}^+_{1\cdot} & \mathbf{0}'_2 & \cdots & \mathbf{0}'_2 \\ \mathbf{0}'_2 & \boldsymbol{s}^+_{2\cdot} & \cdots & \mathbf{0}'_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}'_2 & \mathbf{0}'_2 & \cdots & \boldsymbol{s}^+_{\mathrm{T}\cdot} \end{bmatrix},$$

where $\boldsymbol{s}^+_{t\cdot}$ denotes the tth row of $\mathbf{S}^+$ and $\mathbf{0}_2$ is the 2×1 zero vector.

Now consider $(\mathrm{I}_{\mathrm{T}}\otimes 1_3)\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$. Premultiplication of $\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$ by $(\mathrm{I}_{\mathrm{T}}\otimes 1_3)$ simply amounts to premultiplying the elements of $\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$ by 1_3 so that $(\mathrm{I}_{\mathrm{T}}\otimes 1_3)\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$ can be obtained from $\boldsymbol{\Psi}_{\mathrm{T}}(\mathrm{I}_{\mathrm{T}}\otimes\mathrm{S}^+)$ by substituting $1_3\boldsymbol{s}^+_{t\cdot}$ for $\boldsymbol{s}^+_{t\cdot}$, $t = 1,...,T$, and the 3×2 zero matrix $\mathbf{0}_{3,2}$

for $\mathbf{0}_2'$. Finally, since $\mathbf{J}'$, $\mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+$ and $\mathbf{0}_{3,2}$ are all 3×2 the following sum can be formed

$$(\mathbf{I}_T \otimes \mathbf{J}') - (\mathbf{I}_T \otimes \mathbf{1}_3)\boldsymbol{\Psi}_T(\mathbf{I}_T \otimes \mathbf{S}^+) =$$

$$= \begin{bmatrix} \mathbf{J}' & \mathbf{0}_{3,2} & \cdots & \mathbf{0}_{3,2} \\ \mathbf{0}_{3,2} & \mathbf{J}' & \cdots & \mathbf{0}_{3,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{3,2} & \mathbf{0}_{3,2} & \cdots & \mathbf{J}' \end{bmatrix} - \begin{bmatrix} \mathbf{1}_3 \boldsymbol{s}_{1\cdot}^+ & \mathbf{0}_{3,2} & \cdots & \mathbf{0}_{3,2} \\ \mathbf{0}_{3,2} & \mathbf{1}_3 \boldsymbol{s}_{2\cdot}^+ & \cdots & \mathbf{0}_{3,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{3,2} & \mathbf{0}_{3,2} & \cdots & \mathbf{1}_3 \boldsymbol{s}_{T\cdot}^+ \end{bmatrix}$$

which gives the result.

To prove (ii), note that, by (i), $\text{rank}(\mathbf{A}) = \Sigma_{t=1}^{T} \text{rank}(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+)$, where $\text{rank}(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+) \leq 2$ since $\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+$. is 3×2 for $t = 1,\ldots,T$. It thus has to be proved that $\text{rank}(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+) = 2$ for $t = 1,\ldots,T$. The strategy will be to prove that the first 2 rows of $\mathbf{J}' - \mathbf{1}_n \boldsymbol{s}_{t\cdot}^+$ are linearly independent or, equivalently, that the corresponding submatrix has full row rank. By the definition (23) of $\mathbf{J}$, the first 2 rows of $\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+$ can be written $\mathbf{I}_2 - \mathbf{1}_2 \boldsymbol{s}_{t\cdot}^+$ where $\mathbf{I}_2$ is the identity matrix of order 2 and $\mathbf{1}_2$ is the 2×1 unit vector. As $\det(\mathbf{I}_2 - \mathbf{1}_2 \boldsymbol{s}_{t\cdot}^+) = 1 - \boldsymbol{s}_{t\cdot}^+ \mathbf{1}_2$ [cf. Magnus and Neudecker (1988, p. 25)] the matrix $\mathbf{I}_2 - \mathbf{1}_2 \boldsymbol{s}_{t\cdot}^+$ is non-singular, i.e. has rank equal to 2, as long as $\boldsymbol{s}_{t\cdot}^+ \mathbf{1}_2 \neq 1$. But this follows directly from the definition of $\mathbf{S}^+$; the rows $\boldsymbol{s}_{t\cdot}^+$, $t = 1,\ldots,T$, of $\mathbf{S}^+$ contain the predicted cost shares for the 2 first inputs. Since the sum of *all* the predicted cost shares is identically equal to unity the sum of the 2 first predicted shares must be different from unity. While, theoretically, there is a (remote) possibility that $\boldsymbol{s}_{t\cdot}^+ \mathbf{1}_2$ is exactly equal to one for some t, $t = 1,\ldots,T$, the probability of this event is zero since $\boldsymbol{s}_{t\cdot}^+ \mathbf{1}_2$ is continuous on the interval $[0,1]$. Accordingly, with probability 1 $\text{rank}(\mathbf{I}_2 - \mathbf{1}_2 \boldsymbol{s}_{t\cdot}^+) = 2$ for $t = 1,\ldots,T$, implying that $\text{rank}(\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{t\cdot}^+) = 2$ for $t = 1,\ldots,T$, too. Hence, $\text{rank}(\mathbf{A}) = T \cdot 2$. Q.E.D.

Proof of Lemma 4

The property (i) follows from the fact that since Ω^{-1} is positive definite (due to the assumed positive definiteness of Ω) there exists a non-singular matrix P such that $\Omega^{-1} = PP'$; cf. Magnus and Neudecker (1988, Theorem 23). Accordingly,

$$A(D^{-2} \otimes \Omega^{-1})A' = A(D^{-1} \otimes P)(D^{-1} \otimes P')A' = [A(D^{-1} \otimes P)][A(D^{-1} \otimes P)]' .$$

The rank property (ii) is implied by (i) and well known results concerning ranks [cf. Magnus and Neudecker (op. cit., p. 8)] according to which

$$\begin{aligned} \mathrm{rank}[A(D^{-2} \otimes \Omega^{-1})A'] &= \mathrm{rank}\{[A(D^{-1} \otimes P)][A(D^{-1} \otimes P)]'\} \\ &= \mathrm{rank}[A(D^{-1} \otimes P)] = \mathrm{rank}(A) \end{aligned}$$

where the last equality is a consequence of the fact that $(D^{-1} \otimes P)$ is non-singular [cf. (A.12)] and, hence, must be of full rank. Q.E.D.

Proof of Lemma 5

Consider, first, the matrix product $\Psi_T(W' \otimes I_3)$. This product can be written

$$\Psi_T(W' \otimes I_3) = [E_{11}(3) , E_{22}(3) , E_{33}(3)] \begin{bmatrix} w_{11}I_3 & w_{21}I_3 & \cdots & w_{T1}I_3 \\ w_{12}I_3 & w_{22}I_3 & \cdots & w_{T2}I_3 \\ w_{13}I_3 & w_{23}I_3 & \cdots & w_{T3}I_3 \end{bmatrix}$$

where the matrices $E_{ii}(3)$, $i = 1,2,3$, are 3×3 with the (i,i)'th element equal to unity and all other elements equal to zero.

Premultiplication of $w_{ti}I_3$, $t = 1,...,T$, by $E_{ii}(3)$ yields an 3×3 matrix whose i'th row equals the i'th of $w_{ti}I_3$ while all other rows equal 1×3 zero vectors. Premultiplication of $(W' \otimes I_3)$ by Ψ_T thus yields

$$\Psi_T(W' \otimes I_3) = [\mathrm{diag}(w_{1.}) , ... , \mathrm{diag}(w_{T.})] .$$

Further, since $\mathbf{M} = \mathrm{diag}(\boldsymbol{\mu})$ and $\mathrm{diag}(\boldsymbol{\mu})\mathrm{diag}(\mathbf{w}_{t.}) = \mathrm{diag}(\mu_1 w_{t1}, \ldots, \mu_n w_{tn})$,

$$\mathbf{M}\boldsymbol{\Psi}_T(\mathbf{W}' \otimes \mathbf{I}_n) = [\mathrm{diag}(\boldsymbol{\mu} \odot \mathbf{w}_{1.}), \ldots, \mathrm{diag}(\boldsymbol{\mu} \odot \mathbf{w}_{T.})],$$

where $\odot$ denotes the Hadamard product. Finally, combining this result with Lemma 3, one directly obtains the expression sought. Q.E.D.

<u>Proof of Lemma 6</u>

The matrix considered in the lemma is equal to the following sum

$$(\partial \mathbf{B}^c/\partial \boldsymbol{\kappa})'(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3)\mathbf{A} + (\partial \boldsymbol{g}^\star/\partial \boldsymbol{\kappa})'\boldsymbol{\Psi}_T(\mathbf{I}_T \otimes \mathbf{ZB}')\mathbf{A}.$$

Concerning the first term, note that $\partial \mathbf{B}^c/\partial \boldsymbol{\kappa} = (\partial \mathbf{B}^c/\partial \kappa_1, \ldots, \partial \mathbf{B}^c/\partial \kappa_8)$ where $\partial \mathbf{B}^c/\partial \kappa_h = [(\partial \mathbf{b}_{.1}/\partial \kappa_h)', \ldots, (\partial \mathbf{b}_{.5}/\partial \kappa_h)']'$, $h = 1, \ldots, 8$. Hence

$$\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\kappa}}\right]' = \begin{bmatrix} (\partial \mathbf{b}_{.1}/\partial \kappa_1)' & (\partial \mathbf{b}_{.2}/\partial \kappa_1)' & \cdots & (\partial \mathbf{b}_{.5}/\partial \kappa_1)' \\ (\partial \mathbf{b}_{.1}/\partial \kappa_2)' & (\partial \mathbf{b}_{.2}/\partial \kappa_2)' & \cdots & (\partial \mathbf{b}_{.5}/\partial \kappa_2)' \\ \vdots & \vdots & \ddots & \vdots \\ (\partial \mathbf{b}_{.1}/\partial \kappa_8)' & (\partial \mathbf{b}_{.2}/\partial \kappa_8)' & \cdots & (\partial \mathbf{b}_{.5}/\partial \kappa_8)' \end{bmatrix}. \tag{C.1}$$

Further, using the fact that $\mathbf{G} \equiv \mathrm{diag}(\boldsymbol{g}^\star)$ and, subsequently, Lemma 3

$$(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3)\mathbf{A} =$$

$$= \begin{bmatrix} z_{11} g_1^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{1.}^+) & z_{21} g_2^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{2.}^+) & \cdots & z_{T1} g_T^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{T.}^+) \\ z_{12} g_1^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{1.}^+) & z_{22} g_2^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{2.}^+) & \cdots & z_{T2} g_T^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{T.}^+) \\ \vdots & \vdots & \ddots & \vdots \\ z_{15} g_1^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{1.}^+) & z_{25} g_2^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{2.}^+) & \cdots & z_{T5} g_T^\star (\mathbf{J}' - \mathbf{1}_3 \boldsymbol{s}_{T.}^+) \end{bmatrix}. \tag{C.2}$$

Since the elements in the RHS matrices in (C.1) and (C.2) are 1×3 vectors and 3×2 matrices, respectively, these partitions are conformable. The result of premultiplying $(\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3)\mathbf{A}$ by $(\partial \mathbf{B}^c/\partial \boldsymbol{\kappa})'$ can thus be expressed as a matrix whose elements are 1×2 vectors. Carrying out the multiplication one sees that

$$\left[\frac{\partial \mathbf{B}^c}{\partial \boldsymbol{\kappa}}\right](\mathbf{Z}'\mathbf{G} \otimes \mathbf{I}_3)\mathbf{A} = (\boldsymbol{\pi}_{\mathrm{ht}}) = \left[g_t^\star\left[\sum_{j=1}^{5} z_{tj}(\partial \mathbf{b}_{\cdot j}/\partial \kappa_h)'\right](\mathbf{J}' - \mathbf{1}_3 s^+_{t\cdot})\right]$$
$$= \left[z_{t\cdot} g_t^\star \frac{\partial \mathbf{B}'}{\partial \kappa_h}(\mathbf{J}' - \mathbf{1}_3 s^+_{t\cdot})\right], \tag{C.3}$$

where the indices on the typical 1×2 vector element $\boldsymbol{\pi}_{\mathrm{ht}}$ run from 1 to 8 and from 1 to T, respectively.

Concerning the term $(\partial g^\star/\partial \boldsymbol{\kappa})'\Psi_T(\mathbf{I}_T \otimes \mathbf{ZB}')\mathbf{A}$, consider first the matrix $\Psi_T(\mathbf{I}_T \otimes \mathbf{ZB}')$. Noting that the tth row of $\mathbf{ZB}'$ is equal to $z_{t\cdot}\mathbf{B}'$ one can immediately infer the structure of this matrix from the structure of the matrix $\Psi_T(\mathbf{I}_T \otimes \mathbf{S}^+)$ considered in the proof of Lemma 3. Together with Lemma 3 itself this yields

$$\Psi_T(\mathbf{I}_T \otimes \mathbf{ZB}')\mathbf{A} =$$
$$= \begin{bmatrix} z_{1\cdot}\mathbf{B}'(\mathbf{J}' - \mathbf{1}_3 s^+_{1\cdot}) & \mathbf{0}_2' & \cdots & \mathbf{0}_2' \\ \mathbf{0}_2' & z_{2\cdot}\mathbf{B}'(\mathbf{J}' - \mathbf{1}_3 s^+_{2\cdot}) & \cdots & \mathbf{0}_2' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_2' & \mathbf{0}_2' & \cdots & z_{T\cdot}\mathbf{B}'(\mathbf{J}' - \mathbf{1}_3 s^+_{T\cdot}) \end{bmatrix}.$$

The 1×2 elements of this matrix are conformable with the 1×1 elements of the $8 \times T$ matrix $(\partial g^\star/\partial \boldsymbol{\kappa})' = (\partial g_t^\star/\partial \kappa_h)$. Thus,

$$\left[\frac{\partial g^\star}{\partial \boldsymbol{\kappa}}\right]'\Psi_T(\mathbf{I}_T \otimes \mathbf{ZB}')\mathbf{A} = (\boldsymbol{\phi}_{\mathrm{ht}}) = \left[z_{t\cdot}\frac{\partial g_t^\star}{\partial \kappa_h}\mathbf{B}'(\mathbf{J}' - \mathbf{1}_n s^+_{t\cdot})\right],$$

where the indices on the typical 1×2 vector $\boldsymbol{\phi}_{\mathrm{ht}}$ run from 1 to 8 and from 1 to T, respectively. Accordingly, $\boldsymbol{\pi}_{\mathrm{ht}}$, given by (C.3), and $\boldsymbol{\phi}_{\mathrm{ht}}$ can be summed. Doing so, one obtains the desired result. Q.E.D.

References

Atkinson, S.E. and R. Halvorsen, 1984, Parametric Efficiency Test, Economies of Scale and Input Demand in U.S. Electric Power Generation, *International Economic Review* 25, pp. 647 – 662.

Atkinson, S.E. and R. Halvorsen, 1986, The Relative Efficiency of Public and Private Firms in a Regulated Environment: The Case of U.S. Electric Utilities, *Journal of Public Economics* 29, pp. 281 – 294.

Barten, A.P., 1969, Maximum Likelihood Estimation of a Complete System of Demand Equations, *European Economic Review* 1, pp. 7 – 73.

Berndt, E.R., B.H. Hall, R.E. Hall, and J.A. Hausman, 1974, Estimation and Inference in Nonlinear Structural Models, *Annals of Economic and Social Measurement* 3/4, pp. 653 – 665.

Berndt, E.R. and D.O. Wood, 1975, Technology, Prices, and the Derived Demand for Energy, *Review of Economics and Statistics* 57, pp. 259 – 268.

Diewert, W.E., 1971, An Application of the Shephard Duality Theorem: A Generalized Leontief Cost Function, *Journal of Political Economy* 79, pp. 481 – 507.

Farrell, M.J., 1957, The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society Series A* 120, pp. 253 – 290.

Färe, R. and C.A.K. Lovell, 1978, Measuring the Technical Efficiency of Production, *Journal of Economic Theory* 19, pp. 150 – 162.

Ferrier, G.D. and C.A.K. Lovell, 1990, Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence, *Journal of Econometrics* 46, pp. 229 – 245.

Guilkey, D.K. and C.A.K. Lovell, 1980, On the Flexibility of the Translog Approximation, *International Economic Review* 21, pp. 137 – 147.

Johnston J., 1984, *Econometric Methods*, McGraw-Hill.

Kopp, R.J., 1981, Measuring the Technical Efficiency of Production : A Comment, *Journal of Economic Theory* 24, pp. 450 – 452.

Kopp, R.J. and W.E. Diewert, 1982, The Decomposition of Frontier Cost Function Deviations into Measures of Technical and Allocative Efficiency, *Journal of Econometrics* 19, pp. 319 – 331.

Lau, L.J. and P.A. Yotopoulos, 1971, A Test for Relative Efficiency and an Application to Indian Agriculture, *American Economic Review* 61, pp. 94 – 109.

Lovell, C.A.K. and R.C. Sickles, 1983, Testing Efficiency Hypotheses in Joint Production : A Parametric Approach, *Review of Economics and Statistics* 65, pp. 51 – 58.

Maddala, G.S., 1977, *Econometrics*, New York: McGraw-Hill.

Magnus, J.R., 1988, *Linear Structures*, New York: Charles Griffin & Co.

Magnus, J.R. and H. Neudecker, 1979, The Commutation Matrix: Some Properties and Applications, *Annals of Statistics* 7, pp. 381 – 394.

Magnus, J.R. and H. Neudecker, 1988, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Chichester: John Wiley & Sons.

Pollock, D.S.G., 1979, *The Algebra of Econometrics*, Chichester : John Wiley & Sons.

Rothenberg, T.J., 1971, Identification in Parametric Models, *Econometrica* 39, pp. 577 – 591.

Toda, Y., 1976, Estimation of a Cost Function When the Cost is Not Minimum: The Case of Soviet Manufacturing Industries 1958 – 1971, *Review of Economics and Statistics* 58, pp. 259 – 268.

Toda, Y., 1977, Substitutability and Price Distortion in the Demand for Factors of Production: An Empirical Estimation, *Applied Economics* 9, pp. 203 – 217.

Woodland, A.D., 1979, Stochastic Specification and the Estimation of Share Equations, *Journal of Econometrics* 10, pp. 361 – 383.

CHAPTER IV

An indirect approach to measuring productivity in private services

Scand. J. of Economics 94, Supplement, 229–244, 1992

An Indirect Approach to Measuring Productivity in Private Services

Erik Mellander

Industrial Institute for Economic and Social Research, Stockholm, Sweden

Abstract

Productivity measurement is considered in the context of incomplete output information. Only a *value* measure of output is assumed to be available, which is typical for many service industries. Input markets are assumed to be competitive, while the output market is allowed to be noncompetitive, and potential markups are assumed to be either known or constant. It is shown that if production technologies are homothetic and the elasticities of total costs with respect to output are strictly increasing, the given data are equivalent to complete information, provided the markup is known. If it is not, the results hold *conditional* on the unknown markup.

I. Introduction

A salient feature of the service industry is that, in general, it is very difficult to measure its output. While input data are mostly available, reliable output quantity or output price data usually cannot be found in official statistical sources. Indeed, for some types of services, notably within the public sector, it is not even possible to obtain output value measures. Mellander and Ysander (1990) examined the conclusions that can be drawn about production technology and producer behavior when there is no output information whatsoever. It was shown that for homothetic production technologies, i.e., with the property that the optimal factor mix is independent of the level of production, time series on input prices and input quantities only can be used to study almost all dimensions of the production process by means of a dual approach, using the cost function as the instrument of analysis. In principle, the only aspects that cannot be

* Financial support from the Bank of Sweden Tercentenary Foundation, the Royal Swedish Academy of Sciences and the Sweden–America Foundation is gratefully acknowledged. I have benefited from comments from Thomas Andersson, Ernst Berndt, Rolf Färe, Lars Grönstedt, Sten Nyberg and, in particular, Fabio Schiantarelli and two anonymous referees.

investigated are those affecting input demands neutrally, e.g. (purely) Hicks-neutral technical change and properties relating to returns to scale.[1]

This study extends the analysis to the case where a measure of the value of output is available, which is the typical situation for most kinds of private services. Thus, the analysis is again based on the cost function and presumes that firms are endowed with homothetic production technologies and operate on a competitive input market. The output market, on the other hand, is allowed, to be noncompetitive. It is assumed, however, that in the context of markup pricing, the markup is either known or (approximately) constant. Thus, while there can be a wedge between the marginal cost and the output price, it is assumed that if this wedge is unknown it can be treated parametrically. In order to simplify the analysis, the producer is assumed to maximize profits and attention is confined to static equilibrium models.

The homotheticity assumption and the requirement that, in the absence of *a priori* information, the possible markup be constant might perhaps seem rather restrictive. However, if the traditional route in dealing with the output measurement problem is followed instead and the unknown output is replaced by some proxy variable(s), then it is usually impossible to ascertain the circumstances under which the variations in the proxy(-ies) really mirror the changes in the actual output, and hence whether the results obtained are valid. Here, the conditions under which the method suggested is applicable are completely clear.

It can be argued that the homotheticity assumption is more easily justified in the context of service production than in goods production. Due to the more limited scope for automatization in the service industry, expansion often takes place by setting up additional production units (offices), similar to those already existing, e.g. in the banking industry, travel agencies, etc. As a result, the input proportions change much less than when expansion occurs mainly through additions to the capital stock, as is the case in the manufacturing industry.[2] As regards the constant markup assumption, it should be noted that in the context of productivity measurement, it is quite common to assume not only the input but also the output market to be competitive. Here, the latter assumption is relaxed, albeit in a crude way.

The paper unfolds as follows. In Section II the model is described in terms of the firms' production technologies and the market conditions.

[1] As a reminder, for a homothetic technology returns to scale are determined solely by the level of output and will thus be independent of the input mix.

[2] Of course, input proportions change over time in the service industry as well because of changes in relative factor prices. The claim here is simply that the smaller changes observed for the service industry as compared to the manufacturing industry are due to the fact that *ceteris paribus* expansion affects input proportions much less in the service industry than in the manufacturing industry.

The existence of an equilibrium in the output market is established and sufficient conditions for a unique and stable equilibrium are considered. The key result of the paper is then derived, i.e., a relation between the unknown output variable and the ratio of the value of output over total costs, for which data are assumed to be available. Section III focuses on a decomposition of total factor productivity growth into the effects of (Hicks-)neutral technical change, biased technical change and effects from returns to scale; followed by a general discussion of the estimation of each of these components is also discussed. Implementation of the theoretical results is dealt with in Section IV and some concluding comments are given in Section V.

II. The Model

The basic structure of the model is given by the following three sets of assumptions.

(i) *Technological assumptions*

Firms, indexed by $i = 1, \ldots, m$ where m might be equal to 1, are assumed to produce the (homogeneous) output good by means of (possibly different) homothetic technologies. Given cost minimization, cf. (iii) below, the firm's technology can be characterized by means of the cost function. Due to the homotheticity assumption, the cost function is separable in output, y_i, and the vector $\mathbf{w}$ of input prices, according to

$$C_i = C_i(y_i, \mathbf{w}, t) = f_i(y_i) \cdot C_i(1, \mathbf{w}, t), \qquad \mathrm{i} = 1, \ldots, m, \tag{1}$$

where the time index t represents the state of the technology and $C_i(1, \mathbf{w}, t)$ denotes the cost of producing one unit of output. As regards notation, C_i is only used to denote total costs, to avoid confusion between total and unit cost. Boldface lowercase letters are used to denote vectors. Accordingly, it is assumed here that output can be treated as a scalar. This does not exclude multiple output activities, but it requires the existence of an output aggregate.[3]

The function $f_i(y)$, which is monotonically increasing, determines the scaling properties of the technology.[4] In this and the following section C_i is

[3] This, in turn, amounts to assuming that the optimal output proportions (but not the levels) can be determined without any input information.

[4] In principle, it is conceivable that technological developments might affect the technology's scaling properties, in which case t should also be an argument in the f_i function. For simplicity, I abstract from that possibility here.

assumed to be a regular cost function.[5] In addition, it is taken to be twice differentiable with respect to each of its arguments.

It is also assumed that the elasticity of total costs with respect to output

$$\varepsilon_i = \varepsilon_i(y_i) \equiv \frac{\partial \ln C_i(y_i, w, t)}{\partial \ln y_i} = \frac{y_i \cdot f'_i(y_i)}{f_i(y_i)}, \qquad i = 1, \ldots, m, \tag{2}$$

is monotonically increasing in output, i.e.,

$$\varepsilon'_i(y_i) > 0 \ \forall y_i, \qquad i = 1, \ldots, m. \tag{3}$$

While the cost/output elasticity is very often assumed to be nondecreasing, it is less often assumed to be strictly increasing since this rules out homogeneous technologies and hence, in particular, technologies that are homogeneous of degree 1, i.e., exhibit constant returns to scale.[6] The reason for the strict monotonicity assumption is that later on the existence of a mapping from the cost/output elasticity to the level of output will be exploited; such a mapping exists if, and only if, the function $\varepsilon_i(y_i)$ can be inverted, i.e., if it is strictly monotonic.[7]

It is further assumed that marginal costs are strictly increasing, i.e.

$$f''_i(y_i) > 0 \ \forall y_i, \qquad i = 1, \ldots, m. \tag{4}$$

(ii) *Assumptions about market conditions*

Input markets are assumed to be competitive while the output market is allowed to be noncompetitive. The inverse industry market demand curve

$$p\left(\sum_{i=1}^{n} y_i\right) = p(\mathbf{1}^t \mathbf{y}),$$

where superindex "t" denotes transpose, is assumed to be finite valued, nonnegative, strictly decreasing and twice differentiable.[8] Total industry revenue, i.e., $\mathbf{1}^t \mathbf{y} \cdot p(\mathbf{1}^t \mathbf{y})$, is assumed to be bounded and strictly concave for all $\mathbf{y}$.

---

[5] Regularity conditions can be found in e.g. Diewert (1971). Some of these conditions can be tested statistically; cf. Section IV.

[6] In principle, the only troublesome fact is that constant returns to scale technologies cannot be considered. As technologies that are homogeneous of degree $r \neq 1$ have ever-increasing or ever-decreasing returns to scale and, hence, lack well-defined optimal levels of production, the exclusion of them is not very serious.

[7] This is not to say that homogeneous technologies cannot be analyzed at all — the results in Mellander and Ysander (1990) are also valid for homogeneous technologies. It means, however, that for these technolgies the output value measure yields no extra information in addition to that provided by the input data.

[8] Of course, the argument list of the price function will in general include a number of exogenous shift variables. To simplify the notation, these are suppressed here.

(iii) *Assumptions about information sets and behavior*

All firms are assumed to know the inverse industry market demand curve, their own cost function and the cost functions of all other firms. Given this information, they seek to maximize profits. □

The assumption that production technologies are homothetic implies that the profit maximization problem of firm i can be divided into two separate subproblems. The first is to choose the cost-minimizing factor proportions, which are independent of the scale of production. The second is to choose the optimal level of output.[9] The solution to the first problem is given by $C_i(1, \mathbf{w}, t)$. When solving the second problem, the firm can take $C_i(1, \mathbf{w}, t)$ as given. Accordingly, firm i's maximization problem can be written

$$\max_{y_i} \pi_i = p(\mathbf{1}'\mathbf{y})\cdot y_i - f_i(y_i)\cdot C_i(1, \mathbf{w}, t).$$

It should be noted that assumptions (i) and (ii) imply that the profit function π_i is strictly concave with respect to y_i.[10]

Following Appelbaum (1982), the first-order conditions for profit maximization can be formulated according to

$$p(1-\theta_i\eta) = f_i'(y_i)\cdot C_i(1, \mathbf{w}, t), \qquad i = 1, \dots, m. \tag{5}$$

where

$$\theta_i \equiv (\partial \mathbf{1}'\mathbf{y}/\partial y_i)\cdot(y_i/\mathbf{1}'\mathbf{y}) \tag{6}$$

is the conjectural elasticity of total industry output with respect to the output of firm i and η is the inverse demand elasticity, defined as

$$\eta = -[\partial p(\mathbf{1}'\mathbf{y})/\partial \mathbf{1}'\mathbf{y}]\cdot[\mathbf{1}'\mathbf{y}/p(\mathbf{1}'\mathbf{y})]. \tag{7}$$

According to (6), the firm should set its output such that its marginal cost equals its *perceived* marginal revenue. This formulation of the first-order condition is consistent with a wide range of behavioral modes. For example, under Cournot behavior the conjectural variation $(\partial \mathbf{1}'\mathbf{y}/\partial y_i)$ equals one implying that the conjectural elasticity θ_i reduces to the output

---

[9] It is not possible to separate these two problems for a nonhomothetic technology as the cost-minimizing factor mix will be dependent on the level of production. The fact that I denote the problems "first" and "second", respectively, should not be taken to indicate anything about the order in which they are to be solved; as will be seen later on, it is perfectly possible to begin by considering the second problem.

[10] The strict concavity of total industry revenue with respect to total output implies that $p(\mathbf{1}'\mathbf{y})\cdot y_i$ is strictly concave with respect to y_i. Further, (4) implies that the cost function is strictly convex with respect to y_i or, equivalently, that the negative of the cost function is strictly concave with respect to y_i. Accordingly, π_i is a sum of two strictly concave functions and so must itself be strictly concave.

share of firm i. In the case of perfect competition $\theta_i = 0$ for all i. Further, under pure monopoly and in the case of collusive behavior, the conjectural elasticity will be identically equal to one, in the former case because $y_1 = \mathbf{1}'\mathbf{y}$ and in the latter because $(\partial \mathbf{1}'y/\partial y_i) = \mathbf{1}'y/y_i$ for all i. Although other types of behavior are also conceivable within this framework, only the four types just mentioned are considered here.

In regard to the Cournot oligopoly game, Friedman (1986, pp. 54–56) demonstrates the existence of at least one equilibrium point.[11] Since pure competition can be viewed as a limiting case of the Cournot oligopoly, it follows that there must also be at least one pure competition equilibrium. In the context of pure monopoly the existence of equilibrium is trivial. To the extent that the case of collusive behavior can be treated as a multiplant monopoly operation, i.e., if agreements can really be considered binding, it is clear that there must exist an equilibrium in that case, too.

Conditions for the equilibrium to be unique and stable are considered in an Appendix (available on request) for the simple case where the inverse demand curve is linear. (As is well known, Cournot behavior and collusion yield the same outcome in this case.) It is shown that under this assumption, the equilibrium is unique if there are two firms. For $m = 3$ it is demonstrated that, essentially, the equilibrium is unique if the absolute value of the slope of the demand curve is less than twice the geometric mean of the slopes of the firm's marginal cost curves. For stability it is required, in addition, that the slopes of the firm's marginal cost curves exceed the absolute value of slope of the demand curve if $m = 2$. If $m = 3$ the slopes of the marginal cost curves have to be at least twice the absolute value of the slope of the demand curve.

An equilibrium relation may now be derived between the output level y_i which is presumed to be unknown to the econometrician, and total costs C_i and the value of output $V_i \equiv p \cdot y_i$, for which data are assumed to be available. The first step is to solve (5) for p, yielding

$$p = \kappa_i \cdot f_i'(y_i) \cdot C_i(1, w, t), \qquad i = 1, \dots, m \tag{8}$$

where

$$\kappa_i \equiv \frac{1}{1 - \theta_i \eta}.$$

[11] Assumptions (i)–(iii) combined fulfill Friedman's Conditions 2.1–2.3, with one minor qualification: whereas in Friedman both the inverse demand function and the cost functions are defined over the range $[0, \infty)$, the corresponding range here is assumed to be $[\delta, \infty)$ where δ is some (infinitely) small positive number. The reason for this difference is that the dual cost function is well defined only for strictly positive output levels; cf. Diewert (1971, p. 489). It should also be mentioned that since the intention is to use this model for measuring productivity developments, the Cournot one-shot game has to be regarded as being repeated over time.

According to (8), the output price is given by a markup κ_i over marginal cost. It can easily be shown that $\theta_i\eta$ belongs to the half-open interval $[0, 1[$; cf. Appelbaum (1982, p. 290). Hence, κ_i will be bounded from below by 1, which is its value under perfect competition (since under perfect competition $\theta_i = 0$ for $i = 1, \ldots, m$). The markup will be highest in the contexts of pure monopoly or collusive behavior since in these cases the markup will be equal to $(1-\eta)^{-1}$. In the Cournot case, the markup will lie between these two extremes.

It is assumed that if κ_i is unknown, it can be treated as a constant. Note that, in general, this is *not* the same thing as assuming the price elasticity of demand to be constant; constancy of η is neither a necessary nor a sufficient condition for constancy of κ_i. However, if η is constant then, for κ_i to be constant, θ_i must be constant, too.[12]

The definition of V_i and (8) imply

$$V_i = \kappa_i \cdot f_i'(y_i) \cdot C_i(1, w, t) \cdot y_i, \qquad i = 1, \ldots, m. \tag{9}$$

Hence, by (1) and (2)

$$\frac{V_i}{C_i} = \kappa_i \cdot \varepsilon_i(y_i). \tag{10}$$

Since, according to (3), the function $\varepsilon_i(y_i)$ is invertible, (10) implies that it is possible to express y_i in terms of the ratio V_i/C_i and the markup factor κ_i. This is the key result of the analysis; next we discuss how it can be used in the estimation of total factor productivity growth.[13]

III. On the Estimation of Total Factor Productivity Growth

In this and the following sections the data available to the econometrician are assumed to refer either to a single firm or to an aggregate of firms.[14] Accordingly, the firm index i is dropped in the following.

---

[12] The assumption of a constant κ_i may not be too bad an approximation even if η changes over time, because such changes are likely to be counteracted by changes in θ_i of the opposite sign. For example, take $p(\mathbf{1}'\mathbf{y})$ to be linear. If at a given demand new firms enter — e.g. because costs have been reduced by technical change — then the new equilibrium will be characterized by a higher η than the old one, but also by lower θ_i's, at least given Cournot behavior. Conversely, if at a given industry supply the demand curve shifts outward, then new firms are likely to enter and the new equilibrium will have the same qualitative properties as in the first case.

[13] I was surprised to find that the interesting relation (10) seems to have gone almost unnoticed. However, in a different context, Morrison (1992, p. 55), considers the corresponding result in the monopoly case, i.e., when $\kappa = (1-\eta)^{-1}$.

[14] For the latter case to be meaningful, the existence of a representative firm has to be assumed. As a discussion of aggregation conditions is outside the scope of this paper, suffice it to note that the assumption that all the m firms are identical is (trivially) sufficient for the existence of a representative firm.

The following (time-series) information is assumed to exist. All relevant input data are known, i.e., both the quantities used of the n factors of production, $\boldsymbol{x}=(x_1, \ldots, x_n)$, and the corresponding input prices $\boldsymbol{w}=(w_1, \ldots, w_n)$, and, consequently, total costs $C\equiv \boldsymbol{w}'\boldsymbol{x}$. Regarding the output side, only the value $V\ (\equiv p\cdot y)$ of output is assumed to be known.

The following duality result, due to Ohta (1975), provides a useful decomposition of the growth in total factor productivity (*TFP*). Denote the prodution function corresponding to $C(y, \boldsymbol{w}, t)$ by $\psi(\boldsymbol{x}, t)$. The rate of change in *TFP* can then be written

$$T\hat{F}P \equiv \frac{\partial \ln \psi(\boldsymbol{x}, t)}{\partial t} = \nu \cdot \varepsilon^{-1}, \tag{11}$$

where

$$\nu \equiv -\frac{\partial \ln C(y, \boldsymbol{w}, t)}{\partial t} \tag{12}$$

and ε^{-1} is the inverse of the elasticity of total costs with respect to output, defined in (2). The factor ν is the dual rate of technical change. Thus, if technical change has a positive impact, ν measures the resulting rate of diminution in total costs. The inverse of the cost/output elasticity is the dual rate of return to scale. Returns to scale are increasing if $\varepsilon^{-1}>1$, constant if $\varepsilon^{-1}=1$, and decreasing if $\varepsilon^{-1}<1$.

The dual rate of technical change can be further decomposed into two components corresponding to (Hicks-)neutral technical change and nonneutral, i.e., input specific, technical change. The former is a function of t while the latter depends on both t and $\boldsymbol{w}$. Thus, denoting these functions by g and h,

$$\nu = -\frac{\mathrm{d} \ln g(t)}{\mathrm{d}t} - \frac{\partial \ln h(\boldsymbol{w}, t)}{\partial t}. \tag{12'}$$

We now turn to the problem of estimating the three components in *TFP* growth. The estimation of the dual rate of return to scale, which does not require any assumptions about the functional form of the cost function, is discussed first. Concerning the two components relating to technical change, the one corresponding to nonneutral technical changes is considered briefly, as its estimation is discussed in Mellander and Ysander (1990). Lastly, the direct relation between the presumed output measurement problem and the estimation of (Hicks-)neutral technical change is examined.

The Dual Rate of Return to Scale

Since the dual rate of return to scale is simply the inverse of the cost/output elasticity, (10) yields

$$\varepsilon^{-1} = \kappa^{-1} \cdot \frac{C}{V}, \tag{13}$$

showing that the cost value ratio is proportional to the dual rate of return to scale. Accordingly, if the markup κ is known, the dual rate of return to scale can be computed directly by means of the given data on total costs and the value of output.

However, if κ is not known *a priori*, it is clear that the dual rate of returns to scale effect on total factor productivity can only be measured *conditional* on this unknown constant. That is to say, some kind of sensitivity analysis has to be performed where the consequences of different assumptions about the magnitude of κ are investigated.

The Dual Rate of Nonneutral Technical Change

Estimation of the last term in (12′), $\partial \ln h(w, t)/\partial t$, merely requires input data and the specification of an explicit functional form for the function $h(w, t)$. According to Shephard's lemma

$$S_j = \frac{\partial \ln C}{\partial \ln w_j} = \frac{\partial \ln h(w, t)}{\partial \ln w_j}, \qquad j = 1, \ldots, n, \tag{14}$$

where S_j is the cost share of input j, i.e., $S_j \equiv (w_j x_j / C)$. Thus, the homotheticity assumption makes the cost shares functions of w and t only.

Imposing linear homogeneity of $h(w, t)$ in w, an estimate of $h(w, t)$ can be obtained by simultaneous estimation of $n-1$ of the n share equations.[15] Partial differentiation of this estimate with respect to t then yields an estimate of $\partial \ln h(w, t)/\partial t$.[6]

Before turning to the estimation of the dual rate of Hicks-neutral technical change, it should be said that while the above discussion has shown that the *minimal* requirements for the estimation of $-\partial \ln h(w, t)/\partial t$ are very limited, the efficiency of the parameter estimates might be substantially increased if the cost function is estimated along with the system

[15] As is well known, the system of cost shares is singular and so one of the cost shares has to be left out in the estimation.

[16] Estimation of the system of input cost shares will also yield estimates of the Binswanger (1974) measures of the bias in technical change, and of elasticities of substitution and price elasticities; see Mellander and Ysander (1990).

of cost shares.[17] Hence, the necessity of specifying the cost function completely to enable estimation of the Hicks-neutral component in *TFP* growth, to be discussed next, has the positive side-effect of increasing the precision in the estimate of the function $h(\mathbf{w}, t)$.[18]

The Dual Rate of Neutral Technical Change

As the input cost shares are unaffected by neutral technical change, estimation of the function $g(t)$ requires specification and estimation of the complete cost function. But estimation of the cost function presupposes data on y — or at least data providing information about the variation in y. This is the reason for assumption (3) which, through (10), ascertains that y can be expressed in terms of V, C and κ.

One possibility is to assume that ε is linear in y, i.e.

$$\varepsilon = \beta + \varphi \cdot y, \qquad \varphi > 0,^{19} \tag{15}$$

where the positivity constraint follows from (3). By (10),

$$y = -\frac{\beta}{\varphi} + \frac{1}{\kappa \cdot \varphi} \frac{V}{C}.$$

The specification (15) thus results in y becoming an affine transformation of the V/C ratio. We can go one step further, however, by exploiting the fact that for empirical implementations it is the *variation* in y (rather than its level) that is of interest. The reason is that the explicit cost functions used in empirical applications constitute first- or second-order approximations to the "true" cost function around some point of expansion. Accordingly, what matters are the variations around the expansion point, which means that y (as well as the w_j's and t) are appropriately measured in terms of deviations from this point. The specification

---

[17] Compared to estimating only the system of cost shares, simultaneous estimation of the cost function and the cost shares will increase the efficiency of all the parameter estimates — and hence, in particular, those associated with the function $h(\mathbf{w}, t)$ — for two reasons. First, parametrical constraints between the cost and the share equations will be taken into account explicitly in the latter case. Second, the residual in the cost function and the residuals in the share equations are probably correlated, which can also be taken into account.

[18] Moreover, consideration of the whole cost function in the estimation also makes it possible to test the validity of the restriction that $h(\mathbf{w}, t)$ be linearly homogeneous in $\mathbf{w}$. This is not possible if only the system of cost shares is estimated, since in that case the homogeneity restriction has to be imposed *a priori* to ascertain that the parameters to be estimated are identified.

[19] This type of specification has been discussed by Zellner and Revankar (1969). The corresponding scaling function is given by $f(y) = y^{\beta} \cdot \exp(\varphi \cdot y)$.

(15) may thus be reparameterized according to

$$\varepsilon = \lambda + \varphi \cdot (y - y_0), \qquad \varphi > 0, \tag{15'}$$

where $\lambda \equiv \beta + \varphi \cdot y_0$ and y_0 denotes the point of expansion.

Since the choice of expansion point is arbitrary, we can simply choose the one most convenient to work with. In the following, y_0 is regarded as being equal to the value on y in some "base year", e.g. the mid-point of the observation period. However, y_0 might equally well be set equal to e.g. the observation period mean.

By evaluating (15′) at y and y_0, applying (9) twice, and forming the difference between the results, we obtain

$$y - y_0 = \frac{1}{\kappa \cdot \varphi}[(V/C) - (V/C)_0], \qquad \kappa \geq 1, \quad \varphi > 0, \tag{16}$$

where $(V/C)_0$ denotes the value/cost ratio corresponding to the expansion point, i.e., its base-year value. Thus, by confining our attention to the deviation of y from the expansion point, we arrive at a *proportional* relationship between $[(V/C) - (V/C)_0]$, for which data are assumed to be available, and the unknown output variable.

Of course, there are other specifications of ε which also have the property that ε is monotonically increasing in y. In studies based on e.g. the translog function proposed by Christensen *et al.* (1973), the following formulation is the most common

$$\varepsilon = \alpha + \gamma \cdot (\ln y - \ln y_0) = \alpha + \gamma \cdot \ln(y/y_0) \qquad \gamma > 0. \tag{17}$$

Subjecting (17) to the same operations as those performed on (15′) to obtain (16), we get

$$\ln(y/y_0) = \frac{1}{\kappa \cdot \gamma} [(V/C) - (V/C)_0], \qquad \kappa \geq 1, \quad \gamma > 0. \tag{18}$$

This specification is used in the next section. As a matter of interpretation, note that by taking y_0 to be the value of y in a base year the l.h.s. of (18) becomes (the logarithm of) a quantity index for output.

IV. Implementation by Means of the Translog Cost Function

Since it is desirable to impose few *a priori* restrictions on the substitution possibilities among the factors of production, one should preferably consider flexible functional forms in the specification of an explicit functional form for the general cost function (1). The reason why the translog has been chosen here is that it is convenient to work with and has been shown to provide adequate estimates of quite complex technologies; cf. Guilkey and Lovell (1980). It should be stressed, however, that, in principle, the above results can be implemented by means of any cost function which fulfills the assumptions in Section II.

The translog cost function constitutes a second-order approximation to $\ln C(y, \boldsymbol{w}, t)$ in terms of $\ln y$, $\ln w_1,\ldots,\ln w_n$, and t. Denote the point around which the "true" cost function is expanded by $(\ln y_0, \ln w_{10},\ldots,\ln w_{n0}, t_0)$. The homothetic translog cost function can then be written

$$\ln C = \ln C(y, \boldsymbol{w}, t) = \alpha_0 + \ln f(y) + \ln g(t) + \ln h(\boldsymbol{w}, t) \tag{19}$$

where

$$\ln f(y) \equiv \alpha_y \cdot \ln \underline{y} + \tfrac{1}{2} \cdot \gamma_{yy} \cdot (\ln \underline{y})^2 \tag{20}$$

$$\ln g(t) \equiv \alpha_t \cdot \underline{t} + \tfrac{1}{2} \cdot \gamma_{tt} \cdot \underline{t}^2, \tag{21}$$

$$\ln h(\boldsymbol{w}, t) \equiv \sum_{i=1}^{n} \alpha_i \cdot \ln \underline{w}_i + \frac{1}{2} \cdot \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \cdot \ln \underline{w}_i \ln \underline{w}_j + \sum_{i=1}^{n} \gamma_{it} \cdot \underline{t} \cdot \ln \underline{w}_i \right). \tag{22}$$

and, for notational brevity,

$$\underline{y} \equiv y/y_0, \tag{23a}$$

$$\underline{\boldsymbol{w}} \equiv (\underline{w}_1,\ldots,\underline{w}_n) \equiv (w_1/\omega_{10},\ldots, w_n/w_{n0}), \tag{23b}$$

$$\underline{t} \equiv t - t_0. \tag{23c}$$

Thus, output and the input prices are taken to be measured on index form y_0 and $w_{10},\ldots, w_{n0}$ being the base-year values, i.e., the values at time t_0.

Direct application of (18) yields

$$\ln \underline{y} = \kappa^{-1} \gamma_{yy}^{-1} \cdot \underline{q}, \tag{24}$$

where

$$\underline{q} \equiv (V/C) - (V/C)_0. \tag{25}$$

By means of (24) the cost function can be formulated in terms of V/C, $\boldsymbol{w}$ and t, rather than y, $\boldsymbol{w}$ and t, according to

$$\ln C(V/C, \boldsymbol{w}, t) = \alpha + \ln f^*(V/C) + \ln g(t) + \ln h(\boldsymbol{w}, t), \tag{26}$$

where

$$\ln f^*(V/C) = \alpha_q \cdot \underline{q} + \tfrac{1}{2} \cdot \gamma_{qq} \cdot \underline{q}^2 \tag{27}$$

and

$$\alpha_q = \alpha_y/(\kappa \cdot \gamma_{yy}), \tag{28}$$

$$\gamma_{qq} = 1/(\kappa^2 \cdot \gamma_{yy}).^{20} \tag{29}$$

From (28) and (29) it is clear that for a given κ both α_y and γ_{yy} are identified.

To ensure that the cost/output elasticity implied by the model i.e.

$$\varepsilon = \alpha_y + \gamma_{yy} \cdot \ln y, \tag{30}$$

is consistent with (9) and that the property (3) holds, the following constraints must be imposed on α_q and γ_{qq}

$$0 < \alpha_q. \tag{31}$$

$$\gamma_{qq} = \alpha_q \cdot [(V/C)_0]^{-1}, \tag{32}$$

Together, (28), (29), (31) and (32) yield

$$\alpha_y = \kappa^{-1} \cdot (V/C)_0 \qquad (>0), \tag{33}$$

$$\gamma_{yy} = \kappa^{-2} \cdot \alpha_q^{-1} \cdot (V/C)_0 \qquad (>0). \tag{34}$$

By inserting the expressions for ln y, α_y and γ_{yy} given by (24), (33) and (34) in (30), it can easily be verified that the equality $\varepsilon = \kappa^{-1} \cdot (V/C)$ will always hold, as required by (10). Moreover, (3) holds by positivity of γ_{yy}. Of course, that (3) and (10) hold does not mean that the empirical implementation of the model does not yield any new information about the technology's scaling properties; it will result in estimates of the scaling parameters α_y and γ_{yy} (conditional on κ) and, moreover, it will make it possible to form some idea about the precision in the estimate of the scale elasticity, through the standard error of the parameter α_q of which γ_{yy} is a function. It should also be recalled that it is only by using $\underline{q}$ as an instrument for ln $\underline{y}$ that the function $g(t)$ can be estimated.

Unfortunately, it is not possible to impose *a priori* constraints on the parameters such that (4) is guaranteed to hold. It can be concluded, however, that (4) implies an upper bound on α_q which should be approximately $4 \cdot (V/C)_0$.[21]

Symmetry among the second-order partial derivatives of C with respect to the input prices and linear homogeneity of C in $\boldsymbol{w}$ imply the following constraints

$$\gamma_{jk} = \gamma_{kj}, \qquad \sum_{j=1}^{n} \alpha_j = 1, \qquad \sum_{j=1}^{n} \gamma_{jk} = \sum_{k=1}^{n} \gamma_{jk} = \sum_{j=1}^{n} \gamma_{jt} = 0.$$

[20] If the markup is known, the argument $\underline{q}/\kappa$ is substituted for $\underline{q}$ in the cost function, the parameters α_q and γ_{qq} becoming a_y/γ_{yy} and $\underline{1}/\gamma_{yy}$, respectively.

[21] This conclusion is obtained as follows. By direct calculation it can be shown that (4) holds if and only if $(\varepsilon^2 - \varepsilon + \gamma_{yy}) > 0$, implying that $\gamma_{yy} > 0.25$ is a *sufficient* condition. By inspection of (34) it can be seen that $\gamma_{yy} > 0.25$ translates into the condition $\alpha_q < 4 \cdot (V/C)_0 \cdot \kappa^2$. Since $\kappa \geq 1$, $\alpha_q < 4 \cdot (V/C)_0$ is necessary for $\gamma_{yy} > 0.25$. which in turn is sufficient for (4).

As in the case when output data are available, these restrictions can all be tested.

In empirical applications, the cost function is estimated jointly with $n-1$ of the input cost shares, given by — cf. (14) —

$$S_j = \alpha_j + \sum_{k=1}^{n} \gamma_{jk} \cdot \ln \underline{w}_k + \frac{1}{2} \cdot \gamma_{jt} \cdot \underline{t}.$$

Application of (12)-(12′) to (21) and (22) gives the effect of technical change on the *TFP* growth rate according to

$$v \equiv \frac{\partial \ln C}{\partial t} = -(\alpha_t + \gamma_{tt} \cdot \underline{t}) - \frac{1}{2} \cdot \left(\sum_{j=1}^{n} \gamma_{jt} \cdot \ln \underline{w}_j \right), \tag{35}$$

where the first and second terms correspond to effects from neutral and nonneutral technical change, respectively. Further, by combining (11), (13) and (35)

$$T\hat{F}P = v \cdot \varepsilon^{-1} = -\left(\alpha_t + \gamma_{tt} \cdot \underline{t} + \frac{1}{2} \cdot \sum_{i=1}^{n} \gamma_{it} \cdot \ln \underline{w}_i \right) \cdot \kappa \cdot \frac{C}{V}. \tag{36}$$

Since $T\hat{F}P$ is dependent on κ, it will be necessary to perform a sensitivity analysis on $T\hat{F}P$ with respect to this parameter, unless it is known *a priori.*

Finally, it should be noted that in addition to the productivity measures, the empirical analysis also yields estimates of (the logarithms of) the output quantity and output price indices. (Of course, like the estimates of *TFP* growth, these estimates will be conditional on the markup factor κ.) By means of definition (23a) and results (24) and (34), the log of the output quantity index can be estimated according to

$$\ln(y/y_0) = \underline{q} \cdot [(V/C)_0 \cdot \kappa^{-3} \cdot \alpha_q^{-1}]. \tag{37}$$

Further, the definition

$$\ln(V/V_0) \equiv \ln(py/p_0 y_0) = \ln(p/p_0) + \ln(y/y_0),$$

implies that the log of the output price index can be estimated as

$$\ln(p/p_0) = \ln(V/V_0) - \underline{q} \cdot [(V/C)_0 \cdot \kappa^{-3} \cdot \alpha_q^{-1}]. \tag{38}$$

Since for many service industries proper output quantity and output price indices are not available in the national accounts statistics, (37) and (38) are important by-products of the estimation. For instance, in the

Swedish national accounts, quantity indices for the banking and the insurance industries are obtained by means of the *ad hoc* assumption that average labor productivity increases 2% per annum. The empirical validity of this assumption can be examined by comparing the estimated indices (37) and (38) with the corresponding national accounts indices.

V. Concluding Comments

The problem under consideration concerns the possibilities of characterizing a production process empirically when there is complete input information but the output information is limited to data on the gross value of output — a typical situation in a large part of the private service sector. It is demonstrated that if (i) the technology is homothetic, (ii) output can be treated as a scalar, and (iii) the elasticity of total cost with respect output is strictly increasing in output, then, essentially, the only additional information required for a complete characterization of the production process is the possible difference (in percentage terms) between the marginal cost and the output price, i.e., the potential markup. Since in many cases it is difficult to obtain information about the markup, the analysis proceeds to the case where the price elasticity is unknown but constant. It is shown that in this case the results continue to hold, *conditional* on the unknown markup factor.

The key assumption is (iii); this assumption makes it possible to substitute known variables for the unknown output variable in the cost function. The fact that (iii) is not only sufficient to enable this substitution but also necessary has an important implication: that a value measure of output carries information in excess of that inherent in input data only if the underlying technology is not homogeneous. Thus, that the technology exhibits nonconstant returns to scale is a necessary but not sufficient condition.

As regards productivity measurement, the result is that if the markup is known, the rate of growth in total factor productivity can be estimated with the same precision as if output data were available. If the markup is unknown, the estimated rate of growth in *TFP* will be conditional on the assumption made about the markup. Hence, in applications, it will be necessary to perform a sensitivity analysis where the effect of variations in the markup is assessed. This, however, is quite easy to do; as long as different constant markups are considered, the model does not have to be reestimated when the markup is altered. Moreover, in quite a few empirical applications it should be possible to obtain information at least about the magnitude of the markup, indicating the interval over which the sensitivity analysis should be carried out.

References

Appelbaum, E.: The estimation of the degree of oligopoly power. *Journal of Econometrics 19*, 287–299, 1982.

Binswanger, H. P.: The measurement of technical change biases with many factors of production. *American Economic Review 64*, 964–976, 1974.

Christensen, L. R., Jorgenson, D. W. & Lau, L. J.: Transcendental logarithmic production frontiers. *Review of Economics and Statistics 55*, 28–45, 1973.

Diewert, W. E.: An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy 79*, 481–507, 1971.

Friedman, J.W.: *Game Theory with Applications to Economics.* Oxford University Press, New York, 1986.

Guilkey, D. K. & Lovell, C. A. K.: On the flexibility of the translog approximation. *International Economic Review 21*, 137–147, 1980.

Mellander, E. & Ysander, B.-C.: On the econometric analysis of production when there are no output data. European meeting of the Econometric Society, 1989; revised 1990.

Morrison, C. J.: Markups in U.S. and Japanese manufacturing: A short-run econometric analysis. *Journal of Business & Economic Statistics 10*, 51–63, 1992.

Ohta, M.: A note on the duality between production and cost functions: Rate of returns to scale and rate of technical progress. *Economic Studies Quarterly 25*, 63–65, 1975.

Shephard, R. W.: *Theory of Cost and Production Functions.* Princeton University Press, Princeton, NJ, 1970.

Zellner, A. & Revankar, N. S.: Generalized production functions. *Review of Economic Studies XXXVI*, 241–250, 1969.

Appendix: Conditions for uniqueness and stability of equilibrium

Since the profit functions are assumed to be twice differentiable with respect to the y_j's, Theorem 2.6 in Friedman (1986, p. 45) can be used to formulate conditions under which the equilibrium is unique. According to this theorem, the equilibrium is unique if the symmetric $m \times m$ matrix

$$\mathbf{M} = \mathbf{P} + \mathbf{P}'$$

is negative definite, where $\mathbf{P}$ is the Jacobian matrix of the system of first order derivatives of the profit functions, i.e.

$$\mathbf{P} = \left[\frac{\partial^2 \pi_i}{\partial y_i \partial y_k}\right] .$$

The typical elements of $\mathbf{P}$ are:

$$\frac{\partial^2 \pi_i}{\partial y_i \partial y_k} = p''(1^t\boldsymbol{y}) \cdot y_i + 2 \cdot p'(1^t\boldsymbol{y}) \equiv A_i , \quad i \neq k, \tag{A1}$$

and

$$\frac{\partial^2 \pi_i}{\partial y_i^2} = A_i - f_i''(y_i) \cdot C_i(1, \boldsymbol{w}, t) \equiv A_i - B_i , \tag{A2}$$

implying that

$$\mathbf{M} = \begin{bmatrix} 2(A_1 - B_1) & A_1 + A_2 & \cdots & A_1 + A_m \\ A_2 + A_1 & 2(A_2 - B_2) & \cdots & A_2 + A_m \\ \vdots & \vdots & \ddots & \vdots \\ A_m + A_1 & A_m + A_2 & \cdots & 2(A_m - B_m) \end{bmatrix} .$$

Denote the principal minor subdeterminants of $\mathbf{M}$ by $D_1, D_2, \ldots D_m$ (where, of course $D_m = |\mathbf{M}|$). For $\mathbf{M}$ to be negative definite, the D_j's should alternate in sign, starting with D_1 negative. As the number of firms (i.e. m) grows it becomes exceedingly more difficult to formulate simple conditions which guarantee that the subdeterminants obey these constraints. For this reason, only the cases where $m = 1$, $m = 2$ and $m = 3$ will be considered here.

By the concavity of the profit function the first subdeterminant is always (strictly) negative. The second subdeterminant can be written

$$D_2 = 4[(A_1 - B_1)(A_2 - B_2) - A_1A_2] - (A_1 - A_2)^2 ,$$

which should be positive. A sufficient (but not necessary) condition for $D_2 > 0$ is that the inverse industry market demand curve is linear. Then $A_1 = A_2 = A < 0$ implying that the first term is strictly positive [since the B_i's are strictly positive by (4)], while the second term is zero.

Finally, to simplify the analysis when $m = 3$ only the case with a linear demand curve, i.e. $A_1 = A_2 = A_3 = A$, will be considered. In this case, one obtains, after a number of tedious manipulations, the following expression

$$D_3 = -2 \cdot A^3 + 6\bar{B}_a \cdot A^2 + 18(\bar{B}_g^3 / \bar{B}_h) \cdot A - 6\bar{B}_g^3 ,$$

where $\bar{B}_a$, $\bar{B}_g$, and $\bar{B}_h$ denote the arithmetic, geometric, and harmonic means of B_1, B_2, and B_3, respectively. Further, since $\bar{B}_a \geq \bar{B}_g \geq \bar{B}_h$ (the inequalities being strict unless $B_1 = B_2 = B_3$)

$$D_3 \leq -2 \cdot A^3 + (6 + \epsilon)\bar{B}_g \cdot A^2 + 18\bar{B}_g^2 \cdot A - 6\bar{B}_g^3 , \qquad \text{(A3)}$$

where $\epsilon = (\bar{B}_a - \bar{B}_g) / \bar{B}_g$. It can easily be verified that if $\epsilon \leq 0.5$, which seems like a very reasonable assumption, then the RHS of (A3) will be non-positive for all A such that $-2 \cdot \bar{B}_g \leq A \ (< 0)$. (For higher values on ϵ the lower bound will be closer to zero.) Thus, for the case when $m = 3$ it should be possible to conclude that the equilibrium is unique if the inverse industry market demand curve is linear and (the absolute value of) its slope is less than twice the geometric mean of the slopes of the firm's marginal cost curves.

Stability conditions can be found in Friedman (1977, p. 71). According to these, the equilibrium is stable if

$$A_i - B_i + |\Sigma_{k \neq i} A_k| \ < \ 0 \qquad i = 1,\ldots,m . \qquad \text{(A4)}$$

If the inverse demand curve is linear, then (A4) reduces to

$$(1 - m) \cdot p'(1^t \boldsymbol{y}) - f_i''(y_i) \cdot C_i(1, \boldsymbol{w}, t) \ < \ 0 , \quad i = 1,\ldots,m ,$$

i.e. for $m = 2$ the slope of the marginal cost curve should exceed the absolute value of the slope of the demand curve for each firm. If $m = 3$ the slopes of the marginal cost curves must be more than twice the absolute value of the slope of the demand curve. These conditions are considerably stronger than those required for uniqueness and fulfillment of them implies fulfillment of the uniqueness conditions.

References

Friedman, J.W., 1977, *Oligopoly and the Theory of Games*, Amsterdam: North-Holland.

Friedman, J.W., 1986, *Game Theory with Applications*, New York: Oxford University Press.

IUI dissertations

Researchers who have worked on their degree at IUI and the title of their dissertation. (Doctor's degree in economics unless otherwise stated.)

1944–1954

Folke Kristensson (business administration). *Studier i svenska textila industriers struktur* (Studies in the Structure of Swedish Textile Industries). 1946.

Jan Wallander. *Flykten från skogsbygden* (The Exodus from the Forest Regions). 1948.

Erik Dahmén. *Svensk industriell företagarverksamhet* (Entrepreneurial Activity in Swedish Industry). 1950.

Roland Artle. *Svenskt distributionsväsende* (Swedish Distributive Trades). 1952.

Ragnar Bentzel. *Inkomstfördelningen i Sverige* (The Distribution of Income in Sweden). 1952.

1955–1965

Odd Gulbrandsen. *Strukturomvandlingen i jordbruket* (The Structural Transformation in Agriculture). 1957.

John Ekström (licentiate). *Den textila konsumtionen* (The Consumption of Textiles). 1958.

Göran Albinsson-Bruhner (licentiate). *Svensk verkstadsindustri. Struktur och utvecklingstendenser* (The Engineering Industries in Sweden. Structure and Tendencies of Development). 1961.

Lars Nabseth. *Löneökningars verkningar inom industrin. En studie av anpassningsprocessen inom företaget* (The Effects of Wage Increases in Industry. A Study of the Process of Adaptation within the Firm). 1961.

Erik Höök. *Den offentliga sektorns expansion. En studie av de offentliga civila utgifternas utveckling åren 1913–58* (The Expansion of the Public Sector – A Study of the Development of Public Civilian Expenditures in Sweden 1913–58). 1962.

Bengt G. Rundblad (sociology). *Arbetskraftens rörlighet. En studie av en lokal arbetsmarknad* (The Mobility of Labour. A Study of a Local Labour Market). 1964.

1966–1976

Karl G. Jungenfelt. *Löneandelen och den ekonomiska utvecklingen. En empirisk-teoretisk studie* (Labour's Share and Economic Development). 1966.

Bengt Höglund (written at Lund University). *Modell och observationer. En studie av empirisk anknytning och aggregation för en linjär produktionsmodell* (A Study of Empirical Implementation and Aggregation for a Linear Model). 1966.

Gunnar Eliasson. *Kreditmarknaden och industrins investeringar* (Manufacturing Industry Finance and Short-Run Investment Behaviour). 1967.

Lars Kritz (licentiate, ethnogeography). *Godstransporternas utveckling i Sverige 1950–66 med utblick mot 1980* (Freight Transportation Trends in Sweden 1950–66 and the Outlook for the Seventies). 1968.

Lennart Ohlsson (licentiate). *Utrikeshandeln och den ekonomiska tillväxten i Sverige 1871–1966* (Foreign Trade and Economic Growth in Sweden 1871–1966). 1969.

Villy Bergström (licentiate, written at Uppsala University). *Den ekonomiska politiken i Sverige och dess verkningar* (Economic Policies in Sweden and their Results). 1969.

Lars Lundberg (licentiate, written at Umeå University). *Kapitalbildningen i Sverige 1861–1965* (Capital Formation in Sweden 1861–1965). 1969.

Lars Wohlin. *Skogsindustrins strukturomvandling och expansionsmöjligheter* (Forest-Based Industries: Structural Change and Growth Potentials). 1970.

John Skår. *Produksjon og produktivitet i detaljhandelen* (Production and Productivity in the Retail Sector). 1971.

Bengt Rydén. *Fusioner i svensk industri* (Mergers in Swedish Industry). 1971.

Anders Klevmarken (statistics). *Statistical Methods for the Analysis of Earnings Data with Special Application to Salaries in Swedish Industry*. 1972.

Rolf Rundfelt (business administration). *Reklamens kostnader och bestämningsfaktorer* (Advertising Costs in Sweden – Structure and Determinants). 1973.

Ulf Jakobsson och Göran Normann (partly written at Lund University). *Inkomstbeskattningen i den ekonomiska politiken* (Personal Income Taxation and Economic Policy). 1974.

Göran Eriksson. *Företagens tillväxt och finansiering* (Growth and Financing of the Firm). 1975.

Lennart Ohlsson. *Svensk verkstadsindustris internationella specialisering* (Patterns of Engineering Trade Specialization in Sweden). 1976.

Lars Kritz (ethnogeografy). *Transportpolitiken och lastbilarna* (Transport Policy and the Lorries – A Study of the Effects of Regulation and Deregulation). 1976.

Lars Lundberg (written at Umeå University). *Handelshinder och handelspolitik* (Barriers to Trade and Trade Policy – Studies of the Effects on the Swedish Economy). 1976.

Siv Gustafsson. *Lönebildning och lönestruktur inom den statliga sektorn* (Determination and Structure of Salaries in the Government Sector of Sweden). 1976.

Johan Facht (licentiate). *Emission Control Costs in Swedish Industry*. 1976.

1977–

Hans-Fredrik Samuelsson (business administration). *Utländska direkta investeringar i Sverige* (Foreign Direct Investments in Sweden – An Econometric Analysis). 1977.

Birgitta Swedenborg. *The Multinational Operations of Swedish Firms*. 1979.

Tomas Pousette (licentiate). *Efterfrågan på telefontjänster och telefoner* (Demand for Telephone Services and Telephones). 1976; *Teletjänster – priser och investeringar* (Telephone Services – Prices and Investments). 1979.

Gunnar Du Rietz. *Företagsetableringarna i Sverige under efterkrigstiden* (The Firm Establishments in Sweden in the Post-War Period). 1981.

Richard Murray. *Kommunernas roll i den offentliga sektorn* (The Role of the Local Governments in the Public Sector). 1981.

Jonas Agell (written at Uppsala University). *Tax Reforms and Asset Markets*. 1985.

Kenneth A. Hanson. *Adaptive Economics in Disequlibrium: Essays on Economic Dynamics*. 1986.

Nils Henrik Schager. *Unemployment, Vacancy Durations and Wage Increases: Applications of Markov Processes to Labour Market Dynamics*. 1987.

Joyce Dargay. *Factor Demand in Swedish Manufacturing: Econometric Analyses*. 1988.

Eva Christina Horwitz (licentiate, written partly at the Board of Commerce). *Analys av exportutvecklingen: Metoder och tillämpningar* (Analysis of Export Development: Methods and Applications). 1988.

Christina Hartler (licentiate). *Hushållens fastighetskapital och reala sparande*. (coauthor Jan Södersten.) 1989. *Agricultural Pricing and Growth*. 1991.

Thomas Lindh. *Essays on Expectations in Economic Theory*. 1992.

Eva M. Meyerson (sociology). *The Impact of Ownership Structure and Executive Team Composition on Firm Performance: The Resolution of a Leadership Paradox*. 1992.

Jonas Häckner. *Price and Quality: Essays on product differentiation*.1993.

Sten Nyberg. *Honesty, Vanity and Corporate Equity: Four microeconomic essays*. 1993.

Kent Rune Sjöholm (licentiate). *Hierarchical Modelling of Private Demand in Sweden*. 1993.

Erik Mellander. *Measuring Productivity and Inefficiency Without Quantitative Output Data*. 1993.

IUI Dissertation Series

1. *The Impact of Ownership Structure and Executive Team Composition on Firm Performance: The resolution of a leadership paradox.* Eva M. Meyerson. 1992.
2. *Price and Quality: Essays on product differentiation.* Jonas Häckner. 1993.
3. *Honesty, Vanity and Corporate Equity: Four microeconomic essays.* Sten Nyberg. 1993.
4. *Hierarchical Modelling of Private Demand in Sweden* (licentiate). Kent Rune Sjöholm. 1993.
5. *Measuring Productivity and Inefficiency Without Quantitative Output Data.* Erik Mellander. 1993

ISSN 1102-6057
ISBN 91-7204-426-8