Rendiconti per gli Studi Economici Quantitativi

Editorial Board
Elio Canestrelli (editor in chief)
Marco Corazza
Paola Ferretti
Contents

Measuring the performance of museums: classical and FDH DEA models................................................................. 1
    A. Basso and S. Funari

A characterization for a class of actuarial risk measures.......... 17
    M. Cardin and G. Pacelli

Partial cross ownership and tacit collusion: a comment .......... 27
    C. Corradi and M.C. Corradi

Two symmetric optimal control problems in economics .......... 35
    L. Grosset and B. Viscolani

On the coincidence of system optimum and user equilibrium for a widely used family of cost functions......................... 47
    P. Pellizzari and A. Sorato
Measuring the performance of museums: classical and FDH DEA models

Antonella Basso and Stefania Funari
Dipartimento di Matematica Applicata
Università Ca’ Foscari di Venezia
Dorsoduro 3825/E – 30123 Venezia, Italy
basso@unive.it, funari@unive.it

Abstract. In this paper we use data envelopment analysis to evaluate the performance of museum institutions. Three DEA models are used, which differ in the set of production possibilities taken into consideration: the basic CCR model, the BCC model and the FDH one. An empirical analysis, applied to a set of Italian municipal museums, sheds light on the applicability of the three DEA models considered.

Keywords. Data envelopment analysis, efficiency measure, FDH, museum.

J.E.L. classification: C6, Z1.
M.S.C. classification: 90B50.

1 Introduction

Performance measurement is a relatively new activity in the museum sector. However, this is not an easy task to accomplish since museum activities are often undertaken by non profit organizations, often having a public nature (see for example [15]). For these organizations, we cannot use profit to measure the fulfillment of their objectives.

In this paper we use data envelopment analysis to evaluate the performance of museum institutions. The DEA models used differ in the set of production possibilities taken into consideration: two classical DEA models, namely the well known basic CCR (Charnes, Cooper and Rhodes) model and the BCC (Banker, Charnes and Cooper) one, and the FDH (free disposal hull) model.

These three models proves apt to analyze the relative technical efficiency of a set of museums; first applications of these models to the performance evaluation of museums can be found in [24], [23] and [4].

An empirical analysis, applied to a set of Italian municipal museums located on the major Italian art cities, sheds light on the applicability of the three DEA models considered.

The structure of the paper is the following. Section 2 discusses the issue of the measurement of performance in the museum sector. In Section 3 we present the basic CCR DEA model while in Section 4 we discuss the BCC and FDH models. Finally, the results of the empirical analysis are presented in Section 5.
2 The issue of performance evaluation of museums

It has become customary to measure the performance of museum institutions by means of performance indicators (see for example [21]).

Actually, the problem is first how to define and then how to measure the performance, when this concept refers to special organizations, whose activity is outside the dictates of normal market economy. We may refer not only to cultural organizations but also to those institutions offering educational services (schools, universities), health services (hospitals, clinics) and, more generally, to non profit organizations and public agencies.

One way to analyze performance is to define appropriate ratios which reflect the value of some input in relation to an output or vice versa. The various ratios shed light on a particular aspect of performance measurement; however, in such a way only one aspect of the organization under exam can be considered at a time.

On the other hand, the performance indicators which can be formed are numerous and it is important to select those which are suitable and significant for analyzing how the target organizations behave. Some quantitative techniques are available for this, as noted in [2].

[1], for example, suggests different sets of measurable performance indicators that can be used in the museum sector. He subdivides the indicators according to the various areas of museum’s performance that can be of significant interest. For each area he identifies a suitable set of indicators, each of which is able to describe a particular aspect of the museum’s activity.

For instance, in order to analyze the museum’s performance related to the access function, Ames proposes the following indicators: the attendance trend (ratio between current year’s total attendance and the average of the last three years’ attendance), the capacity utilization (ratio between the total annual attendance and the area accessible to the public), the low income accessibility (ratio between hours per week available for free admission and total hours per week), the minority attendance (ratio between annual minority attendance and total attendance) and the general accessibility (average number of hours open).

In effect, each (partial) ratio focuses on a particular aspect of the museum’s performance. On the other hand, different decision makers (for example museum directors, museum management, fundraisers, government, donors) may have different interests regarding museums’ activities and therefore they are likely to prefer different indicators.

With respect to this, the performance indicator approach has some limitations, which become more marked when such indicators are applied to the evaluation of cultural activities. For an extensive analysis of the ”promises and pitfalls” of museum performance indicators, see [21].

To overcome this drawback we may use an approach based on an economic model of productive efficiency in which the museum organizations are considered as (non profit) enterprises that transform inputs (or resources) into outputs ([23]).
From an economic perspective, the efficiency of a firm is generally evaluated by assuming a maximum profit objective. However, it is difficult to apply an economic model of productive efficiency directly to non profit organizations, as has been noted for example by [17] and [22] for schools and universities, by [26] for organizations of the public sector and by [2] for the non profit sector.

This difficulty is mainly related to the absence of a simple objective function, be it the profit or some other overall quantity to be optimized. Indeed, some organizations may have several conflicting objectives; for example, some goals of the organization management may be in contrast with the aims of the financing bodies. Moreover, it may be difficult to define the various goals precisely.

Other causes of difficulties may lie on the kind of services provided by the organizations, which are often of a public nature and not easily measurable; their value is often difficult to compute since a market price may not even be defined. Right the difficulty in identifying and measuring the goals of museum institutions makes the problem of evaluating their performance particularly difficult. In all these situations the traditional approaches to measure the firm’s performance seem to be inadequate and need to be suitably adjusted, modified or extended.

On the other hand, we may focus on the measurement of the institution’s efficiency. In the literature on performance measurement, the firm’s efficiency analyzes the best possible relationships between inputs and outputs: in this sense, an efficiency index may consider the productivity, by defining the ratio between an output obtained and a resource employed, or the unit cost, by computing the ratio between the cost of inputs and the output value.

Different concepts of efficiency may be relevant; in particular, we may use either the allocative (or price) efficiency, which represents the ability to minimize the production costs of a given output, or the technical efficiency, which represents the ability to maximize the amount of output given the input available or use a smaller quantity of input to achieve the same amount of output ([12]).

A simple way of measuring the technical efficiency of an organization consists of evaluating its ability of maximizing the output, given the inputs or, alternatively, of minimizing the input, given the output (see for example [18]). If we would have a unique output objective, such as profit in profit-oriented companies, and a unique measure of the resources inputed in the production process, we could measure the efficiency of a cultural institutions by computing the ratio between the output and the input:

\[
P = \frac{\text{output}}{\text{input}}.
\] (1)

However, museums usually have a multiple objective structure, or rather have a multiple input and multiple output structure, so that the use of a particular productivity ratio inevitably leads to a partial picture of the museum’s efficiency.

In order to obtain a global efficiency indicator that is able to take into account the complete structure of both inputs and outputs, one may adopt a weighted-sums approach in which the ratio between a weighted sum of outputs and a
weighted sum of inputs is considered:

\[ P = \frac{\sum_{r=1}^{t} u_r y_r}{\sum_{i=1}^{m} v_i x_i}, \]  

where \( y_r \) and \( u_r \) denote the amount of output \( r \) (\( r = 1, \ldots, t \)) provided by the organization and the weight assigned to it, respectively, whereas \( x_i \) and \( v_i \) represent the amount of input \( i \) (\( i = 1, \ldots, m \)) that the organization uses and the weight associated to it, respectively.

Of course, different weighting vectors may lead to different values of the global indicator and therefore to different final judgments about efficiency. The weighted-sums approach is straightforward enough; to make the method work, what is needed is the use of appropriate weighting vectors.

There might be a second way of obtaining an overall evaluation of the performance of a cultural organization. It consists in directly applying the weighted-sums approach to a given set of performance indicators. That is, one can first compute a set of \( n \) performance indexes \( P_k (k = 1, \ldots, n) \), each of them giving a partial picture of the organization’s performance. Then one can define a weighted average of these partial performance indicators

\[ P = \sum_{r=1}^{t} w_r P_k, \]  

where each partial indicator \( P_k \) is weighted by a positive scalar weight \( w_k \) (with \( \sum_{k=1}^{n} w_k = 1 \)). This method can be particularly useful when only partial performance indexes are available, and even when the organization’s goals are not well defined.

The use of global performance indicators in evaluating the performance in the cultural sector is not rare. In this connection we may cite [14], who proposes a model for measuring the efficiency of cultural organizations based on a weighted sum of partial indicators of policy objectives. In this context, the weighting can be considered as “a sort of numerical proxy for the significance that the management of an organization attributes to their strategic choices” ([14]).

In both approaches the task of building a global efficiency index is based on an appropriate weighting process. However, in the analysis of museum performance the weighting process of the various objectives is often very difficult, since the weights should reflect the relative importance given to the various goals by decision makers and this is not easy to define.

Moreover, different decision makers may have different preferences and thus a priori different sets of weights can be considered. Actually museums provide various services, which may be considered as different output objectives: exposition, conservation, research, education, development of special exhibitions and so on (see [20] and [19]). These output objectives are often difficult to define and difficult to measure.

In addition, there could be some difficulties in defining a common set of weights to be used to compare the performance of different museums, as each
museum may choose to organize its resources differently, so that the relative values of the various outputs may legitimately differ among museums. For example, a museum might prefer an educational objective while another could give the research aim a higher priority, so that their weighting structures would be different.

With regard to this, it is possible to compare the technical efficiency of various museums by accepting the difficulty in seeking a common set of weights and trying to find, for each museum, an adequate weighting structure that is not fixed in advance but derives directly from an optimization process. The application of the DEA methodology, which will be briefly presented in the next section, goes in this direction.

3 The DEA approach

The DEA technique has been devised to evaluate the relative efficiency of non profit organizations by considering multiple objectives and multiple inputs simultaneously; see for example [6], [7] and [16].

In particular, the DEA technique enables us to measure the technical efficiency of a set of organizations by analyzing the efficient frontier of the set of production possibilities. The optimal solution of the basic DEA linear programming problem yields the radial measure of technical efficiency as defined in Farrell; on this subject see [25] and [13].

The DEA efficiency measure is defined as a ratio of weighted outputs to weighted inputs where the weights are chosen so that each unit is assigned the most favourable weights (and this means that the weights may be different for the various units).

The most favourable weights are computed by solving an optimization problem for each unit analyzed, by maximizing the efficiency ratio of the unit considered, subject to the constraint that the efficiency ratios of the other units, computed with the same weights, have an upper bound of 1.

Let $j = 1, 2, \ldots, n$ denote the various decision making units, $y_{rj}$, with $r = 1, 2, \ldots, t$, indicate the amount of output $r$ for unit $j$ and $x_{ij}$, with $i = 1, 2, \ldots, m$, represent the amount of input $i$ for unit $j$. The DEA efficiency measure for a decision making unit $j_0$, with $j_0 \in \{1, 2, \ldots, n\}$, can be computed by solving the following linear fractional programming problem

$$
\max_{\{v, u_r\}} \quad h_0 = \frac{\sum_{r=1}^{t} u_r y_{rj_0}}{\sum_{i=1}^{m} v_i x_{ij_0}}
$$

subject to

$$
\sum_{r=1}^{t} \frac{u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \quad j = 1, \ldots, n
$$

$$
u_r \geq \epsilon \quad r = 1, \ldots, t
$$

$$
v_i \geq \epsilon \quad i = 1, \ldots, m,
$$
where \( u_r \) and \( v_i \) are the weights given to output \( r \) and input \( i \), respectively.

\( \epsilon \) is a non archimedean infinitesimal constant that prevents the weights be zero, i.e. a positive constant that is smaller than any positive real number and is such that the product of \( \epsilon \) by any real number remains smaller than any positive real number. This means that \( \epsilon \) is not a real number, since for real numbers the archimedean property holds: for any real number \( a > 0 \) there exists another real number \( a/2 \) such that \( a > a/2 > 0 \).

The treatment of constraints such as (6) and (7) involves non-standard mathematics, but for problem (4)–(7) it is not even necessary to specify the value of \( \epsilon \) explicitly, since a special two-phase procedure allows to handle the mathematical program; on this subject see e.g. [9] and [10].

Actually, constraints (6) and (7) can be equivalently written as strictly positive constraints on the variables \( u_r \) and \( v_i \); on the other hand, they cannot be replaced by simple non negativity constraints on the variables, otherwise, in special cases, the optimal solution of the linear fractional problem could indicate as efficient also inefficient decision making units (see [8]). On the other hand, if the strictly constraints on the variables \( u_r \) and \( v_i \) are replaced by non negativity constraints, a weaker notion of efficiency is obtained, called radial or technical efficiency; see e.g [10].

Problems analogous to (4)–(7) must be solved for each decision making unit for which the efficiency measure is needed.

It can be easily seen that the feasible region of problem (4)–(7) is non empty and that the objective function has an upper bound of 1. On the other hand, it can be seen that in the feasible region defined by constraints (5)–(7) the objective function (4) may sometimes have superior without reaching a maximum. This may happen, in some cases, when the optimal solution of the problem analogous to (4)–(7) in which constraints (6)–(7) are replaced by the non negativity constraints \( u_r, v_r \geq 0 \) has some 0 components. A simple example in which this situation occurs is the following problem: let us consider two decision making units with two inputs and an output and let the values of the inputs be \( x_{11} = 1, x_{21} = 3, x_{12} = 1, x_{22} = 4 \) and the values of the output be \( y_{11} = y_{12} = 2 \). In the feasible region the objective function of this problem has a superior of 1 but it does not have a maximum.

When problem (4)–(7) has optimal solution, then it has an infinite number of optimal solutions, since if \( (v_1, \ldots, v_m, u_1, \ldots, u_t) \) is optimal, then \( \beta(v_1, \ldots, v_m, u_1, \ldots, u_t) \) is also optimal for all \( \beta > 0 \). By defining an equivalence relation that partitions the set of feasible solutions into equivalence classes and selecting a solution from each equivalence class, the fractional problem (4)–(7) can be converted into an equivalent linear programming problem. The representative solution in each equivalence class is usually chosen by imposing either the constraint

\[
\sum_{i=1}^{m} v_i x_{ij0} = 1, \tag{8}
\]
as in the input-oriented models, or the constraint

\[ \sum_{r=1}^{t} u_r y_{rj_0} = 1, \]  

as in the output-oriented models.

The linear program in input-oriented form, called input-oriented CCR (Charnes, Cooper and Rhodes), can be written as follows

\[
\text{max} \quad \sum_{r=1}^{t} u_r y_{rj_0} \\
\text{subject to} \\
\sum_{i=1}^{m} v_i x_{ij_0} = 1 \\
\sum_{r=1}^{t} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \quad j = 1, \ldots, n \\
-u_r \leq -\varepsilon \quad r = 1, \ldots, t \\
-v_i \leq -\varepsilon \quad i = 1, \ldots, m.
\]

When the resulting DEA efficiency measure is equal to 1, unit \( j_0 \) turns out to be efficient, in the sense that it is not dominated by the other decision making units in the set, at least with the most favourable weights. Hence, the DEA measure entails Pareto efficient solutions with the efficient units lying on the efficient frontier.

If unit \( j_0 \) is not efficient, the dual of problem (10)-(14)

\[
\text{min} \quad z_0 - \varepsilon \sum_{r=1}^{t} s^+_r - \varepsilon \sum_{i=1}^{m} s^-_i \\
\text{subject to} \\
x_{ij_0} z_0 - s^-_i - \sum_{j=1}^{n} x_{ij} \lambda_j = 0 \quad i = 1, \ldots, m \\
-s^+_r + \sum_{j=1}^{n} y_{rj} \lambda_j = y_{rj_0} \quad r = 1, \ldots, t \\
\lambda_j \geq 0 \quad j = 1, \ldots, n \\
s^-_i \geq 0 \quad i = 1, \ldots, m \\
s^+_r \geq 0 \quad r = 1, \ldots, t \\
z_0 \quad \text{unconstrained}
\]
allows to identify a composite unit, made up of a convenient linear combination
of the units associated to the strictly positive dual variables $\lambda_j$, which has the
same input and output orientation as unit $j_0$, in the sense that its inputs and
outputs are combined in the same proportions as unit $j_0$, and is efficient with the
same weights (see e.g. [10]). The set of units associated to the strictly positive
dual variables is called peer group.

The coefficients of this linear combination are the dual variables $\lambda_j$, so that
the inputs and outputs of this composite unit are

$$\sum_{j=1}^{n} \lambda_j x_{ij} \quad i = 1, \ldots, m$$

and

$$\sum_{j=1}^{n} \lambda_j y_{rj} \quad r = 1, \ldots, t.$$  \hspace{1cm} (23)

This composite unit obtains the same output levels with less inputs and thus
domimates unit $j_0$.

Let us consider the optimal solution of the dual problem (15)–(21). If the
value of $z_0$ in this optimal solution is equal to 1, then unit $j_0$ is said to exhibit a
weak efficiency. This weak efficiency is also called radial or technical efficiency,
because it is not possible to obtain the same output level by reducing simultane-
ously the level of all inputs; therefore, the input levels cannot be simultaneously
reduced without altering the proportions (the mix) in which they are used (see
e.g. [10]).

If, in addition, in the optimal solution of the dual problem (15)–(21) all the
so called slack variables $s^+_i$ and $s^-_i$ are equal to zero, then we obtain a stronger
definition of efficiency, called Pareto-Koopmans efficiency. In this case, it is not
possible to improve any input or output without worsening some other input or
output ([10]).

On the other hand, if in the optimal solution of the dual problem we have
$z_0 = 1$ but some of the slack variable are strictly positive, then we have a radial or
technical efficiency but at the same time we observe a so called mix inefficiency.
Actually, in this case it would be possible to obtain the same output levels by
altering the input mix in such a way as to reduce the level of at least one input
without having to increase the level of the others.

4 DEA models with constraints on the composite units

Sometimes a convexity constraint is imposed to the composite unit, so that only
convex combinations of units with strictly positive dual variables are considered.
This result is obtained by adding the convexity constraint

$$\sum_{j=1}^{n} \lambda_j = 1$$  \hspace{1cm} (24)
to the set of constraints of the dual problem (16)-(21). The resulting model is called BCC model and was proposed by [3]. It can be seen that the introduction of the convexity constraint (24) has two main consequences (see [3] and [10]). The former is that the set of the feasible production possibilities of the BCC model is the convex hull of the existing decision making units. The latter is that the production frontier is piecewise linear and concave, which leads to a production function which allows the presence of variable returns to scale. This is an important difference with respect to the CCR model, which is based on the assumption of constant returns to scale.

In addition, the BCC model permits to decompose the global efficiency measure found with the CCR model into its technical and scale components. In fact, the BCC efficiency measure turns out to be a pure technical efficiency measure, which is not affected by scale effects. The ratio between the CCR efficiency measure $E_{CCR}$ and the BCC measure $E_{BCC}$ yields the scale efficiency factor $S$. Hence, we have

$$E_{CCR} = E_{BCC} \times S,$$

so that the inefficiency of a decision making unit having $E_{CCR} < 1$ can be due either to a pure technical inefficiency (if the BCC measure $E_{BCC}$ is less than 1) or to a scale inefficiency (when the scale factor $S$ is less than 1, which means that the unit is operating at a non optimal scale).

Another DEA approach, called FDH (Free Disposal Hull), allows composite units which dominate an inefficient decision making unit to be chosen only among the observed units; in this case, linear combinations of units cannot be considered in the efficiency evaluations. This constraint is imposed by requiring that the dual variables $\lambda_j$, with $j = 1, 2, \ldots, n$, which determine the composition of the composite unit, can take only binary 0 or 1 values and sum up to 1.

The FDH approach was first proposed by [11] in the context of the measurement of technical efficiency and was later formalized in the context of DEA approaches by [27].

The FDH efficiency measure is computed by solving the following (dual) mixed integer linear programming problem

$$\min \quad z_0 - \varepsilon \sum_{r=1}^{t} s_r^+ - \varepsilon \sum_{i=1}^{m} s_i^-$$

subject to

$$x_{ij}z_0 - s_i^- - \sum_{j=1}^{n} x_{ij} \lambda_j = 0 \quad i = 1, \ldots, m \quad (27)$$

$$-s_r^+ + \sum_{j=1}^{n} y_{rj} \lambda_j = y_{rj0} \quad r = 1, \ldots, t \quad (28)$$

$$\sum_{j=1}^{n} \lambda_j = 1 \quad (29)$$
Fig. 1. An example of production possibility sets of CCR and FDH DEA models in a problem with two inputs and one output. The two inputs are measured as input units per value of output. The solid line is the production frontier of the CCR model, while the dashed line represents that of the FDH model.

We may notice that the production possibility set of the FDH model is not convex. This is depicted in Figure 1, which shows an example of production possibility sets of CCR and FDH DEA models in a problem with two inputs and one output. While the production possibility set of the CCR model is convex, that of the FDH model is not convex as it is bounded by a step production frontier. In Figure 1 the efficient units in the CCR model are units $U_1$, $U_3$, and $U_4$ while in the FDH model $U_5$ and $U_6$ are also efficient. The ratio between the distance from point $p_2$ (on the CCR efficient frontier) to the origin $O$ and the distance from unit $U_2$ to the origin $O$ gives the CCR efficiency measure of unit $U_2$

$$E_{CCR,U_2} = \frac{\text{dist}(O, p_2)}{\text{dist}(O, U_2)}.$$  \hspace{1cm} (34)

Analogously, the efficiency measure of $U_2$ in the FDH model is given by the ratio between the distance from $h_2$ (on the FDH efficient frontier) to $O$ and the distance from $U_2$ to $O$

$$E_{FDH,U_2} = \frac{\text{dist}(O, h_2)}{\text{dist}(O, U_2)}.$$  \hspace{1cm} (35)
Let us observe that both efficiency measures are radial efficiency measures that keep the proportion of the two inputs constant, so that unit $U_5$ turns out to be efficient in the FDH model though $U_4$ provides the same amount of output with a lower level of input 2 and the same amount of input 1.

In the CCR model the composite unit of the decision making unit $U_2$ is given by point $p_2$, which is a linear combination of units $U_3$ and $U_4$; so \{ $U_3, U_4$ \} is the peer group of $U_2$. In the FDH model the peer group of $U_2$ consists only of unit $U_3$.

Moreover, it can be seen from Figure 1 that units $U_5$ and $U_6$ are not efficient according to the CCR efficiency measure while they reach the maximum FDH efficiency measure.

5 An empirical analysis on Italian data

The DEA models for the measurement of museum efficiency presented in the previous sections have been applied to a few Italian municipal museums.

Previous empirical applications to the analysis of museum performance using a DEA approach can be found in [24], [23] and [4]. In particular, the set of museums taken into consideration in this empirical analysis is the same considered by [4].

A set of Italian municipal museums are considered, which are located in Bologna, Florence and Venice. The museums considered have similar administrative structures (the museums are governed by local authorities) and operate in a similar environment (they are located in major art centers). The input and output data collected are shown in Table 1 and refer to year 1998.

The input variables are the number of workers and the exhibition area. The output data available are the number of visitors paying the full entrance tickets, the number of visitors paying either a reduced price (including eventual special prices or free admittances), the number of temporary exhibitions carried out by the museum in the period considered and the number of other activities (seminars, conferences, research projects etc.) carried out.

In the analysis different levels of aggregation are considered for the output variables. The data available allow to distinguish four different outputs; however, a small number of inputs and outputs is recommended when the number of decision making units is low, in order to avoid a loss in the discriminatory power of a DEA model, due to the fact that a too high percentage of decision making units turn out to be efficient (see for example [5]). Hence, we have also considered a first aggregation of data in which the number of temporary exhibitions and that of other activities are summed up to form a unique output, and a further aggregation in which all the visitors (paying either a full or a reduced ticket) are jointly considered, too.

Table 2 compares the efficiency measures obtained for the CCR, BCC and FDH models with 4, 3 and 2 outputs, respectively, according to the aggregation level considered. Of course, by aggregating some outputs we get an efficiency measure lower than or equal to the efficiency measure obtained without
Table 1. Input and output data for the Italian municipal museums analyzed. The input variables are the number of workers and the exhibition area, the output variables are the number of visitors paying the full entrance tickets, the number of visitors paying a reduced or a special price or admitted free, the number of temporary exhibitions in the period considered and the number of other activities carried out.

<table>
<thead>
<tr>
<th>Museum</th>
<th>City</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Museo Archeologico</td>
<td>22</td>
<td>7680</td>
</tr>
<tr>
<td>M2</td>
<td>Collezioni comunali d’arte</td>
<td>5</td>
<td>2200</td>
</tr>
<tr>
<td>M3</td>
<td>Museo d’arte industriale</td>
<td>3</td>
<td>550</td>
</tr>
<tr>
<td>M4</td>
<td>Museo Civico Medievale</td>
<td>23</td>
<td>1750</td>
</tr>
<tr>
<td>M5</td>
<td>Galleria d’Arte Moderna</td>
<td>19</td>
<td>3808</td>
</tr>
<tr>
<td>M6</td>
<td>Museo Palazzo Vecchio</td>
<td>20</td>
<td>3282</td>
</tr>
<tr>
<td>M7</td>
<td>Museo Firenze Com’era</td>
<td>5</td>
<td>550</td>
</tr>
<tr>
<td>M8</td>
<td>Museo S. Maria Novella</td>
<td>5</td>
<td>758</td>
</tr>
<tr>
<td>M9</td>
<td>Museo Santo Spirito</td>
<td>1</td>
<td>384</td>
</tr>
<tr>
<td>M10</td>
<td>Museo Racc. della Ragione</td>
<td>3</td>
<td>460</td>
</tr>
<tr>
<td>M11</td>
<td>Museo Correr</td>
<td>14</td>
<td>10000</td>
</tr>
<tr>
<td>M12</td>
<td>Ca’Rezzonico</td>
<td>13</td>
<td>3720</td>
</tr>
<tr>
<td>M13</td>
<td>Museo Vetrario</td>
<td>8</td>
<td>5000</td>
</tr>
<tr>
<td>M14</td>
<td>Palazzo Mocenigo</td>
<td>4</td>
<td>1000</td>
</tr>
<tr>
<td>M15</td>
<td>Palazzo Ducale</td>
<td>44</td>
<td>8900</td>
</tr>
</tbody>
</table>

aggregation (formally, by aggregating we introduce a further constraint in the optimization problem). As it can be seen, there is a loss in the discriminatory power of the performance indicator, which affects mainly the FDH model, since a high number of museums are considered efficient. Actually, with this model, according to the disaggregated 4 output formulation, all museums turn out to be efficient, so that it is not possible to rank these museums: all of them stand first on the list. On the other hand, in the 2 output FDH model not all museums are efficient.

If we compare the values of the efficiency measures obtained with the CCR, BCC and FDH models, shown in table 2, we can see that their values are increasing, i.e.

$$E_{CCR} \leq E_{BCC} \leq E_{FDH}. \quad (36)$$

This is clearly due to the fact that BCC adds an additional constraint to the CCR dual problem (constraint (24)) and, analogously, FDH adds a further constraint with respect to the dual BCC model (constraint (30)). Of course, the optimal value of an optimization problem gets worse when additional constraints are added, and this implies that the optimal value of the dual (minimum) problem rises. This explains why the number of efficient museums is always higher in the FDH model and lower in the CCR model.

We can see from Table 2 that two museums are rated as efficient by all the models considered and for any aggregation level: Museum S. Maria Novella (M8) of Florence and Palazzo Ducale in Venice (M15).
Table 2. Efficiency measures obtained for the CCR, BCC and FDH models with 4, 3 and 2 outputs, respectively, according to the aggregation level considered

<table>
<thead>
<tr>
<th>Museum</th>
<th>CCR</th>
<th>BCC</th>
<th>FDH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E_{CCR4}$</td>
<td>$E_{CCR3}$</td>
<td>$E_{CCR2}$</td>
</tr>
<tr>
<td>$M_1$</td>
<td>0.52</td>
<td>0.43</td>
<td>0.15</td>
</tr>
<tr>
<td>$M_2$</td>
<td>1.00</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>$M_3$</td>
<td>0.41</td>
<td>0.41</td>
<td>0.09</td>
</tr>
<tr>
<td>$M_4$</td>
<td>0.80</td>
<td>0.31</td>
<td>0.14</td>
</tr>
<tr>
<td>$M_5$</td>
<td>1.00</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>$M_6$</td>
<td>0.87</td>
<td>0.87</td>
<td>0.71</td>
</tr>
<tr>
<td>$M_7$</td>
<td>0.19</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td>$M_8$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$M_9$</td>
<td>0.28</td>
<td>0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>$M_{10}$</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>$M_{11}$</td>
<td>0.47</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>$M_{12}$</td>
<td>0.51</td>
<td>0.41</td>
<td>0.24</td>
</tr>
<tr>
<td>$M_{13}$</td>
<td>0.50</td>
<td>0.50</td>
<td>0.29</td>
</tr>
<tr>
<td>$M_{14}$</td>
<td>0.44</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>$M_{15}$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

If we focus on the aggregate two-output models, which is more meaningful given the small number of museums compared, we can notice that Museum Santo Spirito (M9) is inefficient according to the CCR model (since $E_{CCR} = 0.14 < 1$), but it is efficient according to the BCC and FDH models (since $E_{BCC} = E_{FDH} = 1$). This means that this museum is globally inefficient, but its inefficiency is due solely to the scale (it is a very small institution) while it exhibits pure technical efficiency. The scale efficiency factor can be easily computed as follows

$$S = \frac{E_{CCR}}{E_{BCC}}. \quad (37)$$

On the other hand, if we analyze the results obtained with the FDH model, we can notice that most museums turn out to be efficient. This means that if dominance relations are tested only for observed decision making units (and not for their linear combinations), few museums turn out to be dominated.

Table 3 presents the peer groups obtained with the CCR, BCC and FDH two-output models. Of course, in the CCR and BCC models the peer group of an efficient museum is given by the museum itself, while for an inefficient unit the peer group is made up by the museums (one or more) that form the composite unit that outperforms the inefficient museum and has a similar input-output orientation.

In the FDH model, by construction, the peer group is made up of a single museum. We may observe that an efficient unit (which has an FDH efficiency measure equal to 1) can sometimes have a peer unit different from itself; this is the case, for example, of museum M2.
Table 3. Peer groups obtained with the CCR, BCC and FDH two-output models

<table>
<thead>
<tr>
<th>Museum</th>
<th>CCR model</th>
<th>BCC model</th>
<th>FDH model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>$M_8, M_{15}$</td>
<td>$M_8, M_9, M_{15}$</td>
<td>$M_1$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$M_8$</td>
<td>$M_8, M_9$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_3$</td>
<td>$M_{15}$</td>
<td>$M_8, M_9$</td>
<td>$M_3$</td>
</tr>
<tr>
<td>$M_4$</td>
<td>$M_8, M_{15}$</td>
<td>$M_8, M_9$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_5$</td>
<td>$M_8$</td>
<td>$M_8$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_6$</td>
<td>$M_8, M_{15}$</td>
<td>$M_8, M_{15}$</td>
<td>$M_6$</td>
</tr>
<tr>
<td>$M_7$</td>
<td>$M_8$</td>
<td>$M_8, M_9$</td>
<td>$M_7$</td>
</tr>
<tr>
<td>$M_8$</td>
<td>$M_8$</td>
<td>$M_8$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_9$</td>
<td>$M_{15}$</td>
<td>$M_9$</td>
<td>$M_9$</td>
</tr>
<tr>
<td>$M_{10}$</td>
<td>$M_8$</td>
<td>$M_8, M_9$</td>
<td>$M_{10}$</td>
</tr>
<tr>
<td>$M_{11}$</td>
<td>$M_8, M_{15}$</td>
<td>$M_8, M_9, M_{15}$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_{12}$</td>
<td>$M_8, M_{15}$</td>
<td>$M_8, M_9, M_{15}$</td>
<td>$M_8$</td>
</tr>
<tr>
<td>$M_{13}$</td>
<td>$M_{15}$</td>
<td>$M_9, M_{15}$</td>
<td>$M_{13}$</td>
</tr>
<tr>
<td>$M_{14}$</td>
<td>$M_{15}$</td>
<td>$M_8, M_9$</td>
<td>$M_{14}$</td>
</tr>
<tr>
<td>$M_{15}$</td>
<td>$M_{15}$</td>
<td>$M_{15}$</td>
<td>$M_{15}$</td>
</tr>
</tbody>
</table>

An example of this kind is graphically illustrated in Figure 1; unit $U_5$ is on the FDH efficient frontier (so it can be considered efficient) but its peer unit is given by unit $U_4$, which allows to obtain the same output level with a lower quantity of input 2. The decision making unit $U_5$ is efficient, according to the FDH performance indicator, because the radial measures ignore slacks in individual inputs, but its peer unit is given by unit $U_4$ because this unit obtains the same level of output by using a lower quantity of input 2.

As for museum $M_2$, it obtains a lower level of outputs than museum $M_8$ (which constitutes its peer group in the FDH model) by using at the same time a higher level of the second input and the same level of the first one. It is a weak form of efficiency, since $M_2$ exhibits a radial or technical efficiency but a mix inefficiency, since some of the slack variables in the optimal solution of the dual problem (26)–(33) are strictly positive. Hence, it could be possible to obtain the same output levels by reducing the level of an input while keeping constant the level of the other one (thus altering the input mix).

References

A characterization for a class of actuarial risk measures

Marta Cardin\textsuperscript{1} and Graziella Pacelli\textsuperscript{2}

1 Dipartimento di Matematica Applicata
Università Ca’ Foscari di Venezia
Dorsoduro 3825/E – 30123 Venezia, Italy
mcardin@unive.it

2 Istituto di Teoria delle Decisioni e Finanza Innovativa (DE.F.IN.)
Università degli Studi di Ancona
Piazza Martelli 8 – 60121 Ancona, Italy
pacelli@posta.econ.unian.it

Abstract. The idea of coherent risk measures has been introduced by Artzner, Delbaen, Eber, and Heath (1999). We consider a particular class of coherent risk measures used in insurance pricing. These measures are obtained by expansion of TVaR measures. We develop an axiomatic description of these measures.

Keywords. Coherent risk measures, insurance pricing, TVaR.

J.E.L. classification: D810.

1 Introduction

In recent years the axiomatic approach to risk measures has been an important and very active subject.

Risk measures can be characterized by axioms that may be different for various applications of actuarial and financial interest.

In this paper we follow an axiomatic approach to study risk measures in an actuarial context.

In actuarial science the premium principles are considered risk measures, see for example [6], [8], and the insurers are also interested to measure the upper tails of distribution functions, see [1]. Risk measures according to the last meaning are considered and studied in [1], [10], [14], [15]. In particular in [1] there is the definition of coherent risk measure which represents a landmark in the following developments, see [11], [14], [15]. On the other hand in [10] a class of premium principles defined by a transform on the decumulative distribution function is considered.

We propose an axiomatic approach based on a minimal set of properties which characterizes an actuarial risk measure likewise in [1], and that permit us
to obtain an interesting representation for a risk measure. The general lines of the paper are those of [9], [13], where different studies are done.

In particular from a result of Greco [7], we show that the considered actuarial risk measures have a Choquet integral representation with respect to a distorted probability. As it is well known distortion risk measures introduced in the actuarial literature by Wang [10] and they belong to an important class of risk measures that include Value at Risk at level $\alpha$, $V_\alpha$, and Tail Value at Risk at level $\alpha$, $TVaR_\alpha$. The features of $V_\alpha$ and $TVaR_\alpha$ are different because different are the corresponding distortion functions, moreover $V_\alpha$ is not coherent while $TVaR_\alpha$ is coherent, [10].

We obtain for all the functional prices which verify the natural following properties (P1) – (P4), see section 2, an integral representation by a concave distortion function; we observe for example that (P4) asks for the functional price $P$ to preserve the stop-loss order and this is an evident request for almost all functional prices. The obtained representation guarantees us to achieve risk measures which are all coherent (see Theorem 4.4) and which are a convex combination of $TVaR_\alpha$, $\alpha \in [0,1]$.

The practical importance in actuarial science and in finance of $TVaR_\alpha$, as measure of the upper tail of a distribution function, is well known and we refer to [4]. We specify that in general $TVaR_\alpha$ is not the same that $CTE_\alpha$, in fact

$$CTE_\alpha(X) = E(X \mid X > V_\alpha) \quad \alpha \in (0,1)$$

where $V_\alpha$, the Value at Risk, denotes the $\alpha$-th quantile of $X$. It’s not possible to represent $CTE_\alpha$ by a distortion risk measure. When random variables are continuous then $TVaR_\alpha$ and $CTE_\alpha$ coincide. The links with other risk measures, the relations and differences between $TVaR_\alpha$, $CTE_\alpha$, are studied in [4].

Our results extend [3] where the case of continuous bounded random variables is considered.

The paper is organized as follow.

In Section 2 we propose and discuss the properties of actuarial risk measures. In Section 3 we define $TVaR_\alpha$ and we present its most important features in connection with the properties of actuarial risk measure of Section 2. Finally in Section 4 we obtain the integral representation result for actuarial risk measures and the characterization as convex combination of $TVaR_\alpha$, $\alpha \in [0,1]$.

## 2 Properties for actuarial risk measures

We consider an insurance contract in a specified time period $[0,T]$. Let $\Omega$ be the state space and $\mathcal{F}$ the event $\sigma$-field at the time $T$. Let $P$ be a probability measure on $\mathcal{F}$. We consider an insurance contract described by a non-negative random bounded variable $X$, $X : \Omega \to \mathbb{R}$ where $X(\omega)$ represents its payoff at time $T$ if state $\omega$ occurs. In actuarial applications a risk is represented by a nonnegative random variable.
We denote by \( F_X \) the distribution function of \( X \) i.e. \( F_X(x) = P(\omega \in \Omega : X(\omega) \leq x), \quad x \in \mathbb{R} \) and we denote by \( S_X \) the survival function corresponding to \( F_X \).

Frequently an insurance contract provides a franchise and then it is interesting to consider the values \( \omega \) such that \( X(\omega) > a \): in this case the contract pays for \( X(\omega) > a \) and nothing otherwise. Then it is useful to consider also the random variable

\[
(X - a)_+ = \max(X(\omega) - a, 0) \tag{2}
\]

Let \( \mathcal{X} \) be a set of nonnegative, bounded random variables such that \( \mathcal{X} \) has the following property:

a) \( aX, \ (X - a)_+, \ (X - (X - a)_+) \in \mathcal{X} \quad \forall X \in \mathcal{X}, \text{ and } a \in [0, +\infty) \).

We observe that the assumption a) easily does not require that \( \mathcal{X} \) is a vector space.

We denote the insurance prices of the contracts of \( \mathcal{X} \) by a functional \( P \) where

\[
P : \mathcal{X} \to \mathbb{R}. \tag{3}
\]

We propose that the insurance functional price \( P \) satisfies the following properties:

(P1) The price, \( P(X) \), of the insurance contract \( X \) depends only on its distribution \( F_X \).

Frequently this hypothesis is assumed in literature, see for example [13]. The property (P1) says that it is not the state of the world to determine the price of a risk, but the probability distribution of \( X \) assigns the price to \( X \). So risks with identical distributions have the same price.

(P2) \( P(X) \leq \sup_{\omega \in \Omega} X(\omega) \quad \text{for all} \quad X \in \mathcal{X} \).

This is just a natural price condition for any customer who wants to underwrite an insurance policy, see [14].

(P3) \( P(X) = P(X - (X - a)_+) + P((X - a)_+) \quad \forall X \in \mathcal{X} \text{ and } a \in [0, +\infty) \).

This condition splits into two comonotonic parts a risk \( X \) (see for example [5]), and permits to identify the part of premium charged for the risk with the reinsurance premium charged by the reinsurer.

(P4) If \( E(X - a)_+ \leq E(Y - a)_+ \quad \forall a \in [0, +\infty) \Rightarrow P(X) \leq P(Y) \quad \forall X, Y \in \mathcal{X} \).
In other words our functional price $P$ respects the stop-loss order.

We remember that stop-loss order considers the weight in the tail of distributions; when other characteristics are equals, stop-loss order select the risk with less heavy tails. About the implications for a functional price $P$ which preserves stop-loss order see for example [12].

Note that it is possible to do some comments.

**Remark 2.1**

i) From a) we have that $X = 0 \in \mathcal{X}$.

ii) Since all the random variables in $\mathcal{X}$ are bounded, from (P2) and from (P3), we have that $P(0) = 0$.

iii) The functional price $P$ assumes its minimum value zero.

iv) Besides, given a subset $B$ of $\mathcal{X}$ then:

$$\inf \{ P(X) : X \in B \} < +\infty \quad \forall B \subseteq \mathcal{X}$$

because $P$ assumes only finite values.

### 3 TailVaR and distortion function

If $X$ is a random variable the quantile reserve at 100th percentile or Value at Risk at level $\alpha$ is

$$V_\alpha(X) = \inf \{ x \in \mathbb{R} \mid F_X(x) \geq \alpha \} \quad \alpha \in (0, 1)$$

It is possible to define also $V_0(X)$ and $V_1(X)$. Since we shall consider bounded random variable $X$, we shall have $V_1(X) = \max X$. For $V_0(X)$ we find $V_0(X) = -\infty$.

A single quantile risk measure of a fixed level $\alpha$ does not give an information about the thickness of the upper tail of the distribution function of $X$, so that other measures are considered.

The Tail Value at Risk at level $\alpha$, $TVaR_\alpha(X)$, is defined as:

$$TVaR_\alpha(X) = \frac{1}{(1 - \alpha)} \int_{\alpha}^{1} V_\alpha(X)d\alpha \quad \alpha \in (0, 1)$$

It is known, see [10], that given a non negative random variable $X$, for any increasing function $g$ with $g(0) = 0$ and $g(1) = 1$. We can define a risk measure

$$H(X) = \int_{0}^{+\infty} g(S_X(t))dt = \int_{0}^{+\infty} (1 - f(F_X(t)))dt$$

and $g$ is called a distortion function and $f$ is called "dual" of $g$ (see [2]).

We observe that the premium in (7) depends only on the distribution function $F_X$ of a random variable $X$.

Both the quantile Value at Risk and TailVar can be obtained with a transformation like in (7) by a special function $g$ (see [11]). When the function $g$ in (7) is concave, we obtain a "concave distortion risk measure".
The quantile Value at Risk is not a concave risk measure, while TailVar is a concave risk measure. In fact, $TVaR_\alpha$ can be obtained by (7) where $f$ is the function defined as follows:

$$f(u) = \begin{cases} 
0 & u < \alpha, \\
\frac{u - \alpha}{(1 - \alpha)} & u \geq \alpha 
\end{cases}$$

(8)

4 Representation theorem

Before deriving the representation theorem, we show some preliminary results.

**Proposition 4.1**

All concave distortion risk measures satisfy the properties ($P1$) − ($P4$).

**Proof**

It follows easily (see [10]).

**Proposition 4.2**

Let $\mathcal{X}$ be a set of non-negative, bounded random variables with property a). If the insurance functional price $P : \mathcal{X} \to \mathbb{R}$ satisfies properties ($P1$) − ($P4$), then $P$ can be represented as the following integral

$$P(X) = \int_{0}^{+\infty} (1 - f(F_X(t)))dt$$

(9)

where $f$ is an increasing function $f : [0, 1] \to [0, 1]$, with $f(0) = 0, f(1) = 1$.

**Proof**

We observe that $\mathcal{X}$ consists of bounded random variables so that

$$P(X) = \lim_{n \to +\infty} P(X - (X - n)_+) \quad \forall X \in \mathcal{X}$$

(10)

I.e. the price of any insurance contract can be obtained by approximation from below.

Moreover

$$P(X) = \lim_{n \to +\infty} P((X - \frac{1}{n})_+) \quad \forall X \in \mathcal{X}$$

(11)

from Remark 2.1 iv) and Proposition 1.2 in [7].

Then we have already observed in Remark 2.1 ii) that $P(0) = 0$.

On the other hand for any $X, Y \in \mathcal{X}$ if $X(\omega) \leq Y(\omega) \forall \omega \in \Omega$ then

$$E(X - a)_+ \leq E(Y - a)_+ \quad \forall a \in [0, +\infty)$$

and from ($P4$) this implies $P(X) \leq P(Y)$: i.e. $P$ ensures the monotonicity property.
So the functional $P$ satisfies the properties (1)-(5) in [7], then from Greco’s Theorem ([7]) we can define two set functions

$$\alpha(A) := \sup\{P(X) | X \in \mathcal{X}, X \leq 1_A\} \quad (12)$$

$$\beta(A) := \inf\{P(Y) | Y \in \mathcal{X}, 1_A \leq Y\} \quad (13)$$

with $A \in \mathcal{F}$ and $1_A$ is the indicator function of set $A$.

Since $P$ is monotone from (P4) then $\alpha, \beta$ are monotone functions and $\alpha \leq \beta$.

Let $\gamma$ be a monotone set function on $\mathcal{F}$ so that $\alpha(A) \leq \gamma(A) \leq \beta(A) \forall A \in \mathcal{F}$, then it follows (see [7]):

$$P(X) = \int_{\Omega} Xd\gamma \quad (14)$$

We observe that generally $\gamma$ is not unique.

Since $P$ satisfies property (P1), then it is possible to write $\gamma$ as a distorted probability, $g \circ P$, by a distortion function $g$ such that

$$P(X) = \int_{\Omega} Xd\gamma = \int_{0}^{\infty} g(S_X(t))dt = \int_{0}^{\infty} (1 - f(F_X(t)))dt \quad (15)$$

and $f$ in (15) is the dual function of $g$ (see [2]).

QED

**Proposition 4.3**

If $f$ is a continuous increasing convex function, defined on $[0, 1]$ then there exists a probability measure $\mu$ on $[0, 1]$ such that

$$f(x) = \int_{0}^{1} \frac{(x - \alpha)_{+}}{(1 - \alpha)} d\mu(\alpha) \quad (16)$$

for $x \in [0, 1]$.

**Proof**

It follows easily, see for example [3].

We give now the representation theorem for the functional $P$ which satisfies properties (P1) – (P4).

**Theorem 4.4**

Let $\mathcal{X}$ be a set of non negative, bounded, random variables with property a).

A functional price $P : \mathcal{X} \to \mathbb{R}$ satisfies properties (P1) – (P4), if and only if there exists a probability measure $m$ on $[0, 1]$ such that:

$$P(X) = \int_{0}^{1} TV aR_{\alpha}(X)dm(\alpha) \quad (17)$$
Proof

Suppose that \( P \) satisfies properties (P1) – (P4), we show that a probability measure exists such that (17) follows.

From Proposition 4.2 we have that
\[
P(X) = \int_0^{+\infty} (1 - f(F_X(t)))dt
\]
(18)
where \( f : [0, 1] \rightarrow [0, 1] \) is an increasing function with \( f(0) = 0 \) and \( f(1) = 1 \) i.e. \( f \) is a distortion function.

Note that \( P \) satisfies the comonotonic additivity property, [5]. Moreover \( P \) verifies (P4) then \( P \) is subadditive, see [12], so the function \( f \) in (18) is a convex function (see for example [2]).

Given an increasing continuous convex function \( h \) on \([0, 1]\), which values zero in zero, from Proposition 4.3 it follows that a probability measure \( m(\alpha) \) exists such that \( h \) can be represented as
\[
h(x) = \int_0^1 \frac{(x - \alpha)_+}{(1 - \alpha)} dm(\alpha)
\]
(19)
where \( \alpha \in [0, 1], h(1) = 1 \), in this way \( h \) is a distortion function, with the property of convexity.

Then, by choosing \( f(x) = h(x) \) in (18), and by interchanging the integrals for the Fubini Theorem, we have:
\[
P(X) = \int_0^{+\infty} (1 - f(F_X(t)))dt = \int_0^{+\infty} \left[ 1 - \int_0^1 \frac{(F_X(t) - \alpha)_+}{(1 - \alpha)} dm(\alpha) \right] dt = \]
\[
= \int_0^{+\infty} dt \int_0^1 \left[ 1 - \frac{(F_X(t) - \alpha)_+}{(1 - \alpha)} \right] dm(\alpha) = \]
\[
= \int_0^1 dm(\alpha) \int_0^{+\infty} dt \left[ 1 - \frac{(F_X(t) - \alpha)_+}{(1 - \alpha)} \right] = \]
\[
= \int_0^1 dm(\alpha) TV aR_\alpha \]
(20)
i.e. the thesis.

Now we suppose that
\[
P(X) = \int_0^1 TV aR_\alpha(X) dm(\alpha)
\]
(21)
and we show that \( P \) satisfies properties (P1)-(P4).

Since \( TV aR_\alpha \) admits an integral representation by a concave distortion \( g_\alpha \), by a backward procedure from (20) we have that also the price functional \( P \) admits an integral representation by a concave distortion function \( g \), so that properties (P1) – (P4) follow from Proposition 4.1.
From the given representation for $P$ follows that the functional price satisfies further meaningful properties.

**Corollary 4.5**

Let be $\mathcal{X}$ a set of non negative, bounded, random variables with property a). Let $P$ be a functional price $P : \mathcal{X} \to \mathbb{R}$ which satisfies properties $(P1) - (P4)$, then:

i) $P$ has the following properties:
- $E[X] \leq P(X) \quad \forall X \in \mathcal{X}$,
- $P(aX + b) = aP(X) + b \quad \forall X \in \mathcal{X}, a \in [0, +\infty) \text{ and } b \in \mathbb{R}$,
- $P(X + Y) = P(X) + P(Y) \quad \text{for comonotonic } X, Y \in \mathcal{X}$,
- $P(X + Y) \leq P(X) + P(Y) \quad \forall X, Y \in \mathcal{X}$.

ii) $P$ is coherent.

**Proof**

From Proposition 4.2 we have an integral representation for $P$ by a distortion function $g$. In the proof of Theorem 4.4 we show that in the integral representation for $P$ the distortion function $g$ is a concave function. Hence all the properties in i) easily follow (see [5]).

The assertion ii) follows according by the definition in [1], (see also [14]).

QED

May be useful to obtain the representation for $P$ by $V_\alpha$

**Corollary 4.6**

Let be $\mathcal{X}$ a set of non negative, bounded, random variables with property a). Let be $P$ a functional price $P : \mathcal{X} \to \mathbb{R}$ which satisfies properties $(P1) - (P4)$. Then there exists probability measure on $[0, 1]$ such that:

$$P(X) = \int_0^1 [V_\alpha(X) + \frac{1}{(1 - \alpha)}ESF_\alpha(X)]dm(\alpha)$$  \quad (22)

where $ESF_\alpha = E[(X - V_\alpha(X))_+]$, $\alpha \in (0, 1)$.

**Proof**

It is known that:

$$TVaR_\alpha(X) = V_\alpha(X) + \frac{1}{(1 - \alpha)}ESF_\alpha(X) \quad \alpha \in (0, 1).$$  \quad (23)

Hence the results follows from Theorem 4.4.

QED
5 Conclusion

In this paper we have obtained for the insurance functional price $P$ which satisfies properties $(P1) - (P4)$ an integral representation by an increasing convex function $f$ with $f(0) = 0$ and $f(1) = 1$. We have shown that it is possible obtain $f$ by probability measures $m(\alpha), \alpha \in [0,1]$ and then it is possible represent $P$ by the integral:

$$P(X) = \int_0^1 TVaR_\alpha(X) dm(\alpha)$$

We point out that in the Choquet integral representation for $P$ we use an integral representation for the function $f$, ”dual” of a distortion $g$. Then the results obtained for the class of insurance functional prices seems interesting both because the class of functionals is determined from few natural properties and these functional prices follow closely linked together to a well known risk measure as $TVaR_\alpha$, $\alpha \in [0,1]$. Moreover we point out that the most important properties for a functional price follow easily from the obtained representation.

References


Partial cross ownership and tacit collusion: a comment

Corrado Corradi\textsuperscript{1} and Marco Claudio Corradi\textsuperscript{2}

\textsuperscript{1} Dipartimento di Matematica per le Scienze Economiche e Sociali
\textit{Atma Mater Studiorum} - Università degli Studi di Bologna
Viale Filopanti 5 – 40126 Bologna, Italy
ccorradi@economia.unibo.it.

\textsuperscript{2} Università degli Studi di Siena
Doctoral student in Law and Economics
wcorradi@libero.it

Abstract. The coordinated competitive effects of silent ownerships in rival firms have been studied recently in the context of an infinitely repeated Bertrand oligopoly model by Gilo and Spiegel (2003). Focusing on their investigation, the purpose of the present note is twofold, to provide a general comparative static result and to examine to what extent their findings are compatible with a more general scenario in which no restriction is placed on the levels of reciprocal ownership stakes.

Keywords. Partial cross ownership, repeated Bertrand oligopoly, tacit collusion.

J.E.L. classification: D43, L41.

1 Introduction

The last several years have witnessed an increased interest in the potential collusive effects of passive cross ownerships among rival firms and the related legal implications, involving, in particular, closer scrutiny policies by competition Authorities, see e.g. Gilo [2], O’Brien and Salop [7], Dubrow [1]. From the standpoint of economic analysis, while the unilateral pricing effects of passive cross ownerships have been investigated in the context of static oligopoly models by several Authors (see e.g. [9] and references cited therein), the coordinated competitive effects have been studied by Malucel [4], in the context of a repeated Cournot game and more recently, by Gilo and Spiegel [3], using the framework of a repeated Bertrand model. Focusing on the latter contribution, the purpose of the present short note is twofold:

a) to provide a formal proof of a general comparative static result, concerning the effect of a change in the level of investment of a firm in any one of the rivals on the ability of firms to engage in tacit coordination;
b) to examine two issues - the transfer of ownership from one rival to another, and the role of passive investments in rival firms by controllers - investigated by the Authors under the restrictive assumption that each firm holds the same ownership stake in each rival, and analyze to what extent their findings can be extended to a more general scenario in which no restriction is placed on the levels of reciprocal investments.

2 Some preliminary notations and results

2.1 The model

Following [3] and using the same notation, we assume that there are \( n \) identical firms competing \( \text{\`a la Bertrand} \) in the market. The firms produce perfect substitutes with the same constant marginal cost; the lower price firm gets the whole market, while the firms share equally the market when they charge the same price. As shown e.g. in Tirole [10], the fully collusive outcome in which all firms charge the monopoly price can be sustained as a subgame perfect equilibrium of the infinitely repeated game if and only if the firms’ intertemporal discount factors exceed the critical level \( \hat{\delta} = 1 - \frac{1}{n} \). In order to examine the impact of partial cross ownership, let \( \beta_j \) be the ownership stake held by the controller of firm \( j \) and let \( a_{ij} \) be firm \( i \)’s ownership stake in firm \( j \), with no limit placed on the number of firms that can have a profit claim on any rival; no limit is placed in principle on the level of individual claims as well, but it is assumed that partial ownership reduces to a purely passive financial investment and confers no right of control. Under these assumptions, the profit of firm \( i \) under collusion is then given by

\[
\pi_i = a_{i1}\pi_1 + a_{i2}\pi_2 + \cdots + a_{i,i-1}\pi_{i-1} + a_{i,i+1}\pi_{i+1} + \cdots + a_{in}\pi_n + \pi^m/n,
\]

for all \( i = 1, \cdots, n \), where \( \pi^m \) denotes the monopoly profit. On the other hand, if the firm \( i \)’s controller deviates from the fully collusive scheme (slightly undercutting the rivals’ price) we have

\[
\pi_i^{d_i} = a_{i1}\pi_1^{d_i} + a_{i2}\pi_2^{d_i} + \cdots + a_{i,i-1}\pi_{i-1}^{d_i} + a_{i,i+1}\pi_{i+1}^{d_i} + \cdots + a_{in}\pi_n^{d_i} + \pi^m;
\]

for all remaining firms we have similar equations with the last term equal to zero. Solving the corresponding systems of linear algebraic equations we obtain \( \pi_i \), \( \pi_i^{d_i} \), for all \( i = 1, \cdots, n \): letting

\[
\hat{\delta}_i = 1 - \frac{\pi_i}{\pi_i^{d_i}},
\]

it is easy to show (see Gilo and Spiegel, cit., Proposition 1) that, with partial cross ownership, the fully collusive scheme can be sustained as a subgame perfect equilibrium of the infinitely repeated game if and only if the intertemporal discount factor exceeds the highest \( \hat{\delta}_i \), corresponding to the industry maverick -
the firm for which the ratio between the collusive profit and the profit following a deviation from the collusive scheme is minimal. The analysis of the effects of partial cross ownership on competition can be thus reduced to the analysis of the links between the coefficients \( a_{ij} \) and the highest \( \hat{\delta}_i \) in the industry.

**Remark 1.** The length of the interval of discount factors for which collusion can be supported is a standard measure of the notion of "ease of collusion".

**Remark 2.** The punishment phase in the Bertrand repeated game is a Nash reversion game.

### 2.2 Some properties of nonnegative matrices

Let us note that since

\[
\beta_j + \sum_{r=1}^{n} a_{rj} \leq 1, \quad j = 1, \cdots, n
\]

(the l.h.s. is the sum of the ownership stakes held by firm \( j \)'s controller and rival firms in each firm \( j \)) the matrix \( A = [a_{ij}] \), with nonnegative entries, and zeroes in the main diagonal, satisfies the Solow conditions

\[
\sum_{r=1}^{n} a_{rj} < 1, \quad j = 1, \cdots, n.
\]

As a consequence, let \( L = I - A = [l_{ij}] \) and denote by \( l^{ij} \) the \((i, j)\)-th element of \( L^{-1} \); then for each \( j = 1, \cdots, n \) we have

\[
l^{ij} > l^{ji} \geq 0, \quad \forall \ i \neq j, \quad i = 1, \cdots, n,
\]

moreover, for any arbitrary positive \( n \)-vector \( z \) the following inequalities hold:

\[
\frac{\sum_{r=1}^{n} z_r l^{ir}}{\sum_{r=1}^{n} z_r l^{jr}} > \frac{l^{ji}}{l^{ij}} \quad \forall \ i \neq j, \quad i, j = 1, \cdots, n,
\]

see e.g. Murata [6]. The inequalities (2), originally proved by Morishima and Nosse [5], will play a crucial role throughout the discussion that follows.

**Remark 1.** In Theorem 16 of Murata, cit., where a proof of the inequalities (2) is reported, the matrix \( A \) is assumed to be indecomposable: however, it can be easily seen that such an assumption does play no role in the proof, and therefore it can be dropped.

### 3 Some applications

Let us first note that we can write

\[
\hat{\delta}_i = 1 - \frac{\pi_i}{\pi_i^d} = \delta - \frac{1}{n} \sum_{r \neq i} l^{ir}/l^{ii}.
\]

We are now ready to consider the following issues.
3.1 Increase in the level of investment in a rival.

Proposition 1 For each $i, h, k$, where $h = k$, we have

$$\frac{\partial \delta_i}{\partial a_{hk}} \leq 0.$$ 

Proof 1 From the formula

$$\frac{\partial l^p q}{\partial l^{uv}} = -p^{u v},$$

that can be obtained immediately by differentiating the identity $L^{-1}L = I$, it follows

$$\frac{\partial \delta_i}{\partial a_{hk}} = \frac{\partial \delta_i}{\partial l_{hk}} \frac{dl_{hk}}{da_{hk}} = \frac{1}{n} \sum_{r \neq i} \frac{l^{ir} l^{ih} l^{ki} - l^{ih} l^{kr} l^{ii}}{(l^i)^2} = \frac{l^{ih}(l^{ki} \sum_{r \neq i} l^{ir} - l^{ii} \sum_{r \neq i} l^{kr})}{(l^i)^2};$$

now, from (2), letting $z = u = (1, 1, \cdots, 1)^T$ we obtain

$$l^i \sum_{r=1}^n l^{ir} < l^{ii} \sum_{r=1}^n l^{ir}, \quad \text{or} \quad l^i \sum_{r \neq i} l^{ir} < l^{ii} \sum_{r \neq i} l^{ir}, \quad (4)$$

whence the result.

Remark 1. From the above property (not proved in [3], but merely illustrated in the case $n = 4$ with a somewhat heavy ad hoc computational development) it immediately follows that an increase of any firm’s stake in any rival, holding all remaining stakes fixed, in general results in a decrease of all $\delta_i$’s and therefore a greater ease of tacit collusion.

Remark 2. In particular, for $k = i$ we have

$$\frac{\partial \delta_i}{\partial a_{hi}} = 0$$

whence we directly obtain the result explicitly proved in [3], that is, $\delta_i$ is invariant w.r.t a change in any of the rivals’ investments in firm $i$, so that if firm $i$ is the industry maverick, then, and only then, tacit coordination is not affected by the change: note that the Proposition above rules out the possibility, considered in (3), that such a change can turn another firm into the industry maverick.

The remaining analysis in (3) is developed under the hypothesis of “full symmetry”, that is, all firms hold exactly the same ownership stakes in rivals, $a_{ij} = \alpha$, for all $i, j$, with $i \neq j$. Now, the question may be raised whether the same conclusions can be directly extended to the general case of an arbitrary partial cross ownership matrix. A partial answer is given in the following discussion.
3.2 Transfer of ownership stake in firm $k$ from firm $j$ to firm $i$

The Authors in [3] show that, starting from the symmetric case, this change in cross ownership configuration always hinders tacit collusion. Now, let us drop the assumption of symmetry, and suppose that firm $i$ buys a stake $x$ in firm $k$ from firm $j$ so that, after the transaction, the matrix $A$ becomes:

$$
\bar{A} = A - \bar{x}e_k^T
$$

where $e_k$ is the $k$-th vector of the canonical basis of $\mathbb{R}^n$ and $\bar{x}$ is the $n \times 1$ vector with $-x$ in the $i$-th entry, $x$ in the $j$-th entry and 0’s in all other entries. The effect of the above perturbation in matrix $A$ can be easily investigated using the Sherman-Morrison-Woodbury’s inversion formula, see e.g. Oshima [8]: if we denote $\bar{L} = I - \bar{A}$, we have

$$
\bar{L}^{-1} = L^{-1} - L^{-1} \bar{x} e_k^T L^{-1} \frac{1}{1 + e_k^T L^{-1} \bar{x}}
$$

so that for all $s = 1, \ldots, n$,

$$
\bar{\delta}_s = 1 - \frac{1}{n} e_k^T L^{-1} u = 1 - \frac{1}{n} e_s^T L^{-1} u - \frac{e_k^T L^{-1} \bar{x} e_k^T L^{-1} u}{1 + e_k^T L^{-1} \bar{x}} =
$$

$$
= 1 - \frac{1}{n} \sum_{r=1}^{n} l^r s - x \left[ (l^k i - l^k j) \sum_{r=1}^{n} l^r + (l^s j - l^s i) \sum_{r=1}^{n} l^r k^r \right]
$$

As a consequence,

$$
\frac{d\bar{\delta}_s}{dx} = \frac{1}{n} \frac{(l^s i - l^s j)}{(\cdots)^2} \left( \sum_{r=1}^{n} l^r s - l^s \sum_{r=1}^{n} l^r k^r \right)
$$

From (5) we immediately obtain

$$
\frac{d\bar{\delta}_k}{dx} = 0,
$$

and, thanks to (1) and (4),

$$
\frac{d\bar{\delta}_j}{dx} > 0.
$$
Now, in the case of symmetry the first term in the numerator in (5), which is positive for \( s = i \), vanishes for all other values of \( s \); this implies that the corresponding derivatives vanish and therefore the critical discount factor above which the fully collusive scheme can be sustained as a subgame perfect equilibrium of the infinitely repeated game increases thanks to (6), so that tacit collusion is always hindered. On the other hand, dropping the assumption of symmetry the sign of \( \ell_{si} - \ell_{sj} \) is indeterminate, so that nothing can be said about the sign of the derivative. As a consequence, we can conclude that in general the transaction will surely hinder tacit collusion only if the seller, \( j \), is the industry maverick.

3.3 Partial reciprocal ownership by controllers

Assuming that the controllers may directly acquire ownership stakes in rival firms, let us denote by \( \beta_{ij} \) the stake that firm \( i \)'s controller has in firm \( j \) (so that \( \beta_{ii} = \beta_i \) denotes his controlling stake in firm \( i \)); then the per period collusive payoff and noncollusive payoff of firm \( i \)'s controller are given respectively by

\[
\sum_{r=1}^{n} \beta_{ir} \pi_r , \quad \sum_{r=1}^{n} \beta_{ir} \pi_{di}, \quad i = 1, 2, \ldots, n
\]

so that in order to compute the critical discount factor above which the fully collusive scheme can be sustained as a subgame perfect equilibrium of the infinitely repeated game we have to consider the quantities

\[
\hat{\delta}_i^c = 1 - \frac{\sum_{r=1}^{n} \beta_{ir} \pi_r}{\sum_{r=1}^{n} \beta_{ir} \pi_{di}}, \quad (7)
\]

Remark 1. It is clear that (7) reduces to (3) whenever \( \beta_{ir} = 0 \) for all \( r \neq i \), while if \( \beta_{ir} \neq 0 \) for some \( r \neq i \) we have \( \hat{\delta}_i > \hat{\delta}_i^c \); indeed,

\[
\hat{\delta}_i - \hat{\delta}_i^c = \frac{1}{n} \sum_{r \neq i} \frac{\beta_{ir}}{\beta_{ii}} \left( \ell^{ii} \sum_{s \neq i} \ell^{is} - \ell^{ri} \sum_{s \neq i} \ell^{is} \right) \quad \ell^{ii} \left( \ell^{ii} + \sum_{r \neq i} \frac{\beta_{ir}}{\beta_{ii}} \ell^{ri} \right)
\]

is positive, in virtue of (1) and (4). This implies that, in general, when a controller of a firm invests in some of the firm’s competitors the anticompetitive effect is stronger than in a case in which only the firm itself invests in its competitors, since the range of discount factors for which tacit collusion can be sustained turns out to be expanded.
Remark 2. The critical discount factor depends on the firm for which the ratio
\[ \frac{\sum_{r=1}^{n} \beta_{ir} \sum_{s \neq i}^{l_{rs}}}{\sum_{r=1}^{n} \beta_{ir} l_{ri}} \]
is minimal. We note that while in the case of full symmetry one can simply consider the ratios
\[ \frac{\sum_{r \neq i}^{\beta_{ir}}}{\beta_{ii}} \]
involving only the \( i \)-th controller’s stakes in rivals relative to his stake in his own firm, in general, via the coefficients \( l_{ij} \), the critical discount factor depends on the entire matrix \( [a_{ij}] \) of firm’s ownership stakes in rivals.

Remark 3. We have
\[ \frac{\partial \delta_{c}^i}{\partial \beta_{ih}} = \frac{1}{n} \sum_{r=1}^{n} \beta_{ir} \left( \frac{l_{hi} \sum_{s \neq i}^{l_{rs}} - l_{ri} \sum_{s \neq i}^{l_{hs}}}{\left( \sum_{r=1}^{n} \beta_{ir} l_{ri} \right)^2} \right) \]
so that, in particular, \( \frac{\partial \delta_{c}^i}{\partial \beta_{ii}} > 0 \), thanks to (4): this provides a general formalization to the conventional wisdom that a reduction of a controller’s stake in his own firm (while retaining control over the firm itself) weakens the controller’s incentive to deviate from the collusive scheme, exactly as found in [3] under the assumption of full symmetry. Unlike this case, however, in general the sign of the derivative for any \( h \neq i \) is undefined and consequently from the model as such we can not get any conclusive evidence about the direction of the effect of an increment of a controller’s stake in a rival.

The above results provide a sufficient support to the following...

4 Concluding remarks

The assumption of full simmetry, introduced in [3] "to facilitate the analysis", indeed yields a considerable simplification in the mathematical treatment of the various issues, and allows to obtain some insight about several aspects of the theme of interest. However, it should be pointed out that:

a) on the one hand, a general result on the effects of a change in the cross ownership matrix such as the above Proposition 1 can be proved without resorting to any simmetry assumption;
b) on the other hand, the validity of the conclusions is to be confined to the effects of changes from a configuration of identical ownership stakes, and in general it can not be extended *sic et simpliciter* to an arbitrary *existing* (perhaps more "realistic") configuration: indeed, complexity increases as we move from a simple symmetric example to more complicated structures and the existence, direction and size of effects appear to depend crucially on the precise pattern of ownership.

**Acknowledgements**

The authors wish to thank professor Anna Torriero for introducing [6] to them, and the referees for valuable comments.

**References**

Two symmetric optimal control problems in economics *

Luca Grosset and Bruno Viscolani

Dipartimento di Matematica Pura ed Applicata
Via Belzoni 7, I – 35131 Padova, Italy
viscolani@math.unipd.it

Abstract. Two optimal control problems arising in Economics are analyzed in order to elicit some information about other problems, symmetric with them. Different problems, based on the same economic model, can have very similar sets of necessary conditions. Such a situation occurs when considering different objectives for the same economic system. We prove that such conditions can be reduced to a controllability problem and that they concern the characterization of the Pareto efficient solutions of a multiobjective problem.

Keywords. Optimal control, multiobjective programming, Pareto optimality.

J.E.L. classification: C61.
M.S.C. classification: 49N90, 90C29.

1 Symmetries between problems

A lot of models arising in Economics are studied using optimal control theory. In this paper we want to analyze, from a different point of view, two classical problems: Ramsey’s model [5] and consumption versus investment model [6, p.78]. In both these models total utility is to be maximized, under the condition that a fixed final level for a state variable (e.g. the final value of the investment) is reached at least. Otherwise, it is interesting, from an economic point of view, to study also the (symmetric) problem, in which the state variable is to be maximized under the condition that the total utility reaches at least a fixed final level. In fact, both problems are quite similar and we can prove that the necessary conditions can be reduced to the same controllability problem. We might also formulate a third, two-objective problem, whose Pareto efficient solutions are characterized by the same necessary conditions. The result is essentially an application of the constraint method [2, pp.115–127] for generating noninferior solutions of a multiobjective problem.

The idea of analyzing symmetries among economical problems is well known in Consumer Theory, where the Marshall and Hicks problems are studied in a

* Supported by MIUR-Italy and University of Padua
symmetric way using the nonlinear programming theory. The solutions one can
gain give a deeper understanding of both problems and the link between results
permits an interesting economical analysis [7, 111-167], [1]. Some general results
on the connections between the solutions of a non-linear programming problem
and its reciprocal are given in [8]. One can find there, in particular, the discuss-
on of sufficient conditions to assure that a primal problem and its reciprocal
have the same solution, in a finite dimensional setting. Here we aim at prov-
ing similar results as those in [8] for two couples of symmetric optimal control
problems, concerning two classical economic models, and at reading symmetry
results in terms of a special Pareto efficient frontier. To the best of our knowl-
edge, such symmetric approach is not frequently encountered in the economical
models studied with the optimal control theory. In the Calculus of Variations
the presence of symmetries between different problems is a well known subject
[3, p.143] and, recently, such properties have been used to study a marketing
problem [4]. This paper is meant to follow the same direction and show that one
can reach a deeper knowledge of an economic model by considering symmetries
among different problems.

The paper is organized as follows: in Section 2 we introduce the consumption
versus investment model, we define the symmetric problem and we show that a
controllability problem is the core of the necessary conditions for both problems.
In Section 3 the same analysis is repeated for the Ramsey’s model.

2 Consumption versus investment

A country receives a continuous cash flow \( f(t) > 0 \) as aid during the program-
ing interval \([0, T]\). Let \( U \) be a utility function \((U(0) = 0, U'(0) = +\infty, U'> 0, \)
\( \lim_{x \to +\infty} U'(x) = 0, U'' < 0, U \in C^2) \) and let us denote by \( u(t) \in [0, 1] \) the
part of the aid which is allocated to consumption; the decision represented by
the control function \( u(t) \) produces a utility flow \( U(f(t)u(t)) \). If we define a
state variable \( x_2(t) \) that represents the total utility obtained during the interval
\([0, t]\), then its evolution is described by the equation \( \dot{x}_2(t) = U(f(t)u(t)) \). Let
\( x_1(t) \) be the level of infrastructure at the time \( t \) in the country and let \((1 - u(t))\nbe the part of the aid which is allocated to investment. The investment produces
an increment in the level of infrastructure of the country, which is described by
\( \dot{x}_1(t) = f(t)(1 - u(t)) \). The standard approach to this problem is to require that
the government maximizes the total utility and reaches at least a fixed level of
infrastructure at the end of the programming interval. On the other hand, also a
different economical viewpoint may be interesting: the government might decide
to maximize the final level of the infrastructure under the condition of reaching
a fixed utility value. We want to study these problems in parallel and prove that
the relevant necessary conditions lead to a controllability problem.

Let \( \alpha_T \) be the required minimum final level of the infrastructure and \( \alpha_0 \) be
the initial level for the same state variable. The standard consumption versus
investment problem can be formulated as follows (we will refer to this problem
Symmetric optimal control problems in economics

as the consumption problem):

\[
\max_{u \in L^1([0,T];[0,1])} x_2(T),
\]

s.t. \[
\begin{align*}
\dot{x}_1(t) &= f(t)(1 - u(t)), \\
\dot{x}_2(t) &= U(f(t)u(t)), \\
x_1(0) &= \alpha_0, \quad x_1(T) \geq \alpha_T, \\
x_2(0) &= 0, \quad x_2(T) \in \mathbb{R}.
\end{align*}
\] (1)

Assuming the symmetric point of view, we consider that the country wants to reach a minimal level \(\beta_T\) of utility (fixed \textit{a priori} by the government) and it wants to maximize the final level of the infrastructure, starting from the initial level \(\alpha_0\). Therefore, the investment problem can be defined as follows:

\[
\max_{u \in L^1([0,T];[0,1])} x_1(T),
\]

s.t. \[
\begin{align*}
\dot{x}_1(t) &= f(t)(1 - u(t)), \\
\dot{x}_2(t) &= U(f(t)u(t)), \\
x_1(0) &= \alpha_0, \quad x_1(T) \in \mathbb{R}, \\
x_2(0) &= 0, \quad x_2(T) \geq \beta_T.
\end{align*}
\] (2)

2.1 Necessary conditions

The Hamiltonian function for both problems (1) and (2) is the same:

\[
H(x_1, x_2, p_1, p_2, u, t) = p_1 f(t)(1 - u) + p_2 U(f(t)u).
\]

First of all, we notice that

\[
\begin{align*}
H_u(x_1, x_2, p_1, p_2, u, t) &= -p_1 f(t) + p_2 f(t) U'(f(t)u), \\
H_{uu}(x_1, x_2, p_1, p_2, u, t) &= p_2 f^2(t) U''(f(t)u),
\end{align*}
\]

hence the sign of the adjoint variable \(p_2\) is connected with the concavity/convexity of the Hamiltonian function. The adjoint equations for both problems are the following:

\[
\begin{align*}
\dot{p}_1(t) &= 0, \\
\dot{p}_2(t) &= 0.
\end{align*}
\]

The transversality conditions are different and can be written for the consumption and the investment problem respectively as:

\[
\begin{align*}
\begin{cases}
p_1(T) = \gamma_1, \\
p_2(T) = p_0 + \gamma_2, \\
\gamma_1 \geq 0, \\
\gamma_1 (x_1(T) - \alpha_T) = 0, \\
\gamma_2 = 0, \\
\end{cases}
\quad \begin{cases}
p_1(T) = p_0 + \gamma_1, \\
p_2(T) = \gamma_2, \\
\gamma_1 = 0, \\
\gamma_2 \geq 0, \\
\gamma_2 (x_2(T) - \beta_T) = 0.
\end{cases}
\end{align*}
\]
2.2 Limit instances \((p_0 = 0)\)

In this subsection we look for a solution with \(p_0 = 0\), as the necessary conditions require that \((p_0, \gamma_1, \gamma_2) \neq 0\) and \(p_0 \in \{0, 1\}\).

**Consumption problem** \((1)\)

If \(p_0 = 0\) and \(\gamma_2 = 0\) then \(\gamma_1 > 0\) must hold and therefore, by the transversality conditions, we have that \(x_1(T) = \alpha_T\). From the adjoint equations we obtain that \(p_1(t) = \gamma_1 > 0\) and \(p_2(t) \equiv 0\), hence the Hamiltonian function becomes

\[ H(x_1, x_2, p_1, p_2, u, t) = \gamma_1 f(t) (1 - u), \]

and is maximized w.r.t. the variable \(u \in [0, 1]\) if and only if \(u = 0\). There exists a unique aid flow allocation that satisfies the maximum Hamiltonian condition, that of devoting all the aid to investment: such a solution is optimal if and only if it drives the infrastructure level to the final value \(x_1(T) = \alpha_T\). This occurs if and only if \(\alpha_T = \alpha_{\text{max}}\), where

\[ \alpha_{\text{max}} = \alpha_0 + \int_0^T f(t) \, dt \]

is the maximum reachable infrastructure value.

**Investment problem** \((2)\)

If \(p_0 = 0\) and \(\gamma_1 = 0\) then \(\gamma_2 > 0\) and therefore, by the transversality conditions, we get that \(x_2(T) = \beta_T\). The solutions of the adjoint equations are \(p_2(t) \equiv \gamma_2 > 0\) and \(p_1(t) \equiv 0\), hence the Hamiltonian function

\[ H(x_1, x_2, p_1, p_2, u, t) = \gamma_2 U(f(t) u) \]

is maximized w.r.t. the variable \(u \in [0, 1]\) if and only if \(u = 1\). Symmetrically to the first problem, we find a unique aid flow allocation and it consists in spending all the aid in consumption: it is optimal if and only if it drives the total utility to the final value \(x_2(T) = \beta_T\). This occurs if and only if \(\beta_T = \beta_{\text{max}}\), where

\[ \beta_{\text{max}} = \int_0^T U(f(t)) \, dt \]

is the maximum reachable immediate utility value.

2.3 Degenerate instances

If \(p_0 = 1\), then the condition \((p_0, \gamma_1, \gamma_2) \neq 0\) is satisfied and the transversality conditions read, for problem \((1)\) and problem \((2)\) respectively, as

\[
\begin{align*}
\begin{cases}
   p_1(T) = \gamma_1, \\
   p_2(T) = 1, \\
   \gamma_1 \geq 0, \\
   \gamma_1 (x_1(T) - \alpha_T) = 0,
\end{cases} & \quad \begin{cases}
   p_1(T) = 1, \\
   p_2(T) = \gamma_2, \\
   \gamma_2 \geq 0, \\
   \gamma_2 (x_2(T) - \beta_T) = 0.
\end{cases}
\end{align*}
\]
We consider first some particular choices of the parameters that turn the problems into trivial ones.

**Consumption problem** (1)
If \( \gamma_1 = 0 \), then from the adjoint equations we get \( p_1(t) \equiv 0 \) and \( p_2(t) \equiv 1 \), therefore the Hamiltonian function is

\[
H(x_1, x_2, p_1, p_2, u, t) = U(f(t)u),
\]
and is maximized w.r.t. the \( u \in [0,1] \) by choosing \( u = 1 \). The state function associated with the control \( u(t) \equiv 1 \) satisfies the transversality condition if and only if \( \alpha_T \leq \alpha_0 \). If all the aid is spent in consumption, then the infrastructure level remains constant and the final constraint \( x_1(T) \geq \alpha_T \) is only satisfied when \( \alpha_0 \geq \alpha_T \).

**Investment problem** (2)
Symmetrically, if \( \gamma_2 = 0 \), then we obtain \( p_1(t) \equiv 1 \) and \( p_2(t) \equiv 0 \); therefore maximizing the Hamiltonian

\[
H(x_1, x_2, p_1, p_2, u, t) = f(t)(1 - u),
\]
w.r.t. the \( u \in [0,1] \), leads to the control \( u \equiv 0 \). The associated state function satisfies the transversality condition if and only if \( \beta_T \leq 0 \). If all the aid is invested, then the immediate utility does not increase and the final constraint \( x_2(T) \geq \beta_T \) is not satisfied, unless \( 0 \geq \beta_T \), because \( U(0) = 0 \) and \( U' > 0 \).

**2.4 The controllability problem**
After excluding the previous instances, in which there was no trade off between consumption and investment, we obtain the following form for the transversality conditions of the problems (1) and (2):

\[
\begin{cases}
  p_1(T) = \gamma_1 > 0, \\
  p_2(T) = 1, \\
  x_1(T) = \alpha_T,
\end{cases}
\quad
\begin{cases}
  p_1(T) = 1, \\
  p_2(T) = \gamma_2 > 0, \\
  x_2(T) = \beta_T.
\end{cases}
\]

The adjoint functions \( p_1(t) \) and \( p_2(t) \) are strictly positive constants and such is also their ratio \( \pi = p_1(t)/p_2(t) \in (0, +\infty) \). Hence the Hamiltonian function is strictly concave in \( u \) and it is maximized in \( \mathbb{R} \) at one point such that the first derivative w.r.t. \( u \) vanishes:

\[
U'(f(t)u) = p_1(t)/p_2(t) = \pi.
\]

Now \( U' \) has an inverse function \( \phi : (0, +\infty) \to (0, +\infty) \) so that the above maximum condition is equivalent to \( u = \phi(\pi)/f(t) \). Hence we obtain the control function

\[
u^\pi(t) = \min \left\{ 1, \frac{\phi(\pi)}{f(t)} \right\}.
\]
Now both problems are reduced to the same controllability problem. The government has to fix the constant $\pi \in (0, +\infty)$ in order to drive the system of ODEs
\[
\begin{align*}
\dot{x}_1(t) &= f(t) (1 - u^\pi(t)) , \\
\dot{x}_2(t) &= U(f(t) u^\pi(t)) ,
\end{align*}
\]
from the starting point
\[
\begin{align*}
x_1(0) &= \alpha_0 , \\
x_2(0) &= 0 ,
\end{align*}
\]
to a point in either final surface $S_1 = \{\alpha_T\} \times \mathbb{R}$, for the consumption problem (1), or $S_2 = \mathbb{R} \times \{\beta_T\}$, for the investment problem (2).

Now, let us define $\bar{\pi} = \phi^{-1}(\max f) = U'(\max f)$, which is the minimum observable marginal utility, then $\bar{\pi} > 0$ and

- if $\pi \leq \bar{\pi}$, we have the $u^\pi(t) \equiv 1$ and its associated solution is such that $x_1(T) = \alpha_0$, $x_2(T) = \beta_{\max}$;

- if $\pi \to +\infty$, the control $u^\pi(t)$ tends to $u^\infty(t) \equiv 0$ and $x_1(T) \to \alpha_{\max}$, $x_2(T) \to 0$.

Moreover, the map
\[
\Xi : [\bar{\pi}, +\infty) \longrightarrow [\alpha_0, \alpha_{\max}) \times (0, \beta_{\max})
\]
\[
\pi \mapsto \left( \alpha_0 + \int_0^T [f(t) - \phi(\pi)]^+ dt , \int_0^T U(\min \{f(t), \phi(\pi)\}) dt \right)
\]
is continuous, its first component $\xi_1(\pi)$ is strictly increasing, whereas its second component $\xi_2(\pi)$ is strictly decreasing. Hence $\Xi$ is a 1-1 function. Given a point $\Xi(\pi) = (\xi_1(\pi), \xi_2(\pi))$ on its image, we can state that

- $\xi_1(\pi)$ is the optimal value of $x_1(T)$ in the investment problem (2) with $\beta_T = \xi_2(\pi)$;

- $\xi_2(\pi)$ is the optimal value of $x_2(T)$ in the consumption problem (1) with $\alpha_T = \xi_1(\pi)$.

### 2.5 Pareto efficient solutions

In fact we might think of a third situation, in which the country government has not fixed a priori any infrastructure nor any utility value threshold, but aims at maximizing the two-dimensional objective $(x_1(T), x_2(T))$. From this viewpoint, the country government is interested in the Pareto efficient solutions with respect to the criteria infrastructure value, represented by $x_1(T)$ and immediate utility, represented by $x_2(T)$. Now, solving the controllability problem is equivalent to determining the Pareto efficient solutions.
Let us define the functions
\[ V_2 : [\alpha_0, \alpha_{\text{max}}] \rightarrow \mathbb{R}, \quad V_1 : [0, \beta_{\text{max}}] \rightarrow \mathbb{R}, \]
\[ \alpha_T \mapsto V_2(\alpha_T) = x_2^*(T), \quad \beta_T \mapsto V_1(\beta_T) = x_1^*(T), \]
as the optimal values of the objective functionals of the consumption problem, \( V_2(\alpha_T) \), which we call the immediate utility of infrastructure value, and of the investment problem, \( V_1(\beta_T) \), which we call the infrastructure value of immediate utility.

We notice that \( V_2(\alpha_T) \) is the final value of the state function \( x_2(t) \), associated with the control \( u^{\pi}(t) \), where \( \pi = \pi_1(\alpha_T) \) satisfies the condition
\[ \int_0^T [f(t) - \phi(\pi)]^+ \, dt = \alpha_T - \alpha_0. \]
Hence \( \pi_1(\alpha_T) \) is a continuous and strictly decreasing function of \( \alpha_T \) and thus also \( V_2(\alpha_T) \) is continuous and strictly decreasing.

On the other hand, \( V_1(\beta_T) \) is the final value of the state function \( x_1(t) \), associated with the control \( u^{\pi}(t) \), where \( \pi = \pi_2(\beta_T) \) satisfies the condition
\[ \int_0^T U(\min\{f(t), \phi(\pi)\}) \, dt = \beta_T. \]
Hence \( \pi_2(\beta_T) \) is a continuous and strictly decreasing function of \( \beta_T \) and thus also \( V_1(\beta_T) \) is continuous and strictly decreasing.

From the definition of the map \( \Xi \) in the previous Section, we notice that
- if \( \pi = \pi_1(\alpha_T) \), then
  \[ \xi_2(\pi) = V_2(\alpha_T), \quad \xi_1(\pi) = \alpha_T; \]
- if \( \pi = \pi_2(\beta_T) \), then
  \[ \xi_1(\pi) = V_1(\beta_T), \quad \xi_2(\pi) = \beta_T. \]

Hence, for all \( \pi \in [\bar{\pi}, +\infty) \), we have that
\[ \xi_1(\pi) = V_1(\xi_2(\pi)), \quad \xi_2(\pi) = V_2(\xi_1(\pi)), \]
i.e. \( V_1 = V_2^{-1} \). In other words, we can state the following identity results for the optimal controls of both problems 1 and 2:
- if \( u^*(t) \) is an optimal control of the consumption problem, then it is optimal also for the investment problem, provided that \( \beta_T = V_2(\alpha_T) \);
- if \( u^*(t) \) is an optimal control of the investment problem, then it is optimal also for the consumption problem, provided that \( \alpha_T = V_1(\beta_T) \).
3 Ramsey’s model

The economic growth theory, which is based on the pioneering work of Ramsey [5], is a cornerstone of economic modeling. Let us define:

- $x_1(t)$, $x_2(t)$ the state functions, i.e. the capital value at time $t$ and the total utility produced during the interval $[0, t]$;
- $u(t)$ the control function, i.e. the consumption flow at time $t$;
- $f(x_1)$ the production function, strictly positive, strictly increasing and concave, that represents the production flow associated with the value $x_1$ of the capital stock;
- $U(c)$ the utility function, strictly increasing and strictly concave, that describes the utility flow associated with the rate $c$ of consumption ($U(0) = 0$, $U'(0) = +\infty$, $U' > 0$, $\lim_{c \to +\infty} U'(c) = 0$, $U'' < 0$, $U \in C^2$);
- $\rho \geq 0$ the discount factor.

Therefore the standard Ramsey’s problem is formulated as follows:

$$\max_{u \in L^1([0,T];[0,\infty))} x_2(T) ,$$

s.t.

$$\begin{align*}
\dot{x}_1(t) &= f(x_1(t)) - u(t) , \\
\dot{x}_2(t) &= e^{-\rho t} U(u(t)) , \\
x_1(0) &= \alpha_0 , & x_1(T) &\geq \alpha_T , \\
x_2(0) &= 0 , & x_2(T) &\in \mathbb{R} .
\end{align*}$$

(3)

The decision maker controls the consumption and wants to maximize the final total utility, subject to a lower bound $\alpha_T$ on the capital stock level at the end of the programming interval ($\alpha_0$ is the capital level at time $t = 0$).

Assuming the symmetric point of view, the decision maker has to reach a minimal utility level $\beta_T$ (a priori fixed) and wants to maximize the final capital stock level. Therefore, the symmetric Ramsey’s problem is

$$\max_{u \in L^1([0,T];[0,\infty))} x_1(T) ,$$

s.t.

$$\begin{align*}
\dot{x}_1(t) &= f(x_1(t)) - u(t) , \\
\dot{x}_2(t) &= e^{-\rho t} U(u(t)) , \\
x_1(0) &= \alpha_0 , & x_1(T) &\in \mathbb{R} , \\
x_2(0) &= 0 , & x_2(T) &\geq \beta_T .
\end{align*}$$

(4)

Our aim is to reduce both problems to a suitable controllability problem, as we have done in the case of the consumptions versus investment model.

3.1 Necessary conditions

The Hamiltonian function for both problems is

$$H(x_1, x_2, p_1, p_2, u, t) = p_1 (f(x_1) - u) + p_2 e^{-\rho t} U(u)$$

and its first and second derivatives w.r.t. $u$ are

$$H_u = -p_1 + p_2 e^{-\rho t} U'(u) , \quad H_{uu} = p_2 e^{-\rho t} U''(u) .$$
We obtain the adjoint equations

\[
\begin{cases}
\dot{p}_1 (t) = -p_1 (t)f' (x_1 (t)), \\
\dot{p}_2 (t) = 0,
\end{cases}
\]

and the transversality conditions for the problems (3) and (4) respectively

\[
\begin{cases}
p_1 (T) = \gamma_1, \\
p_2 (T) = p_0 + \gamma_2, \\
\gamma_1 \geq 0, \\
\gamma_1 (x_1 (T) - \alpha_T) = 0, \\
\gamma_2 = 0,
\end{cases}
\quad \begin{cases}
p_1 (T) = p_0 + \gamma_1, \\
p_2 (T) = \gamma_2, \\
\gamma_1 = 0, \\
\gamma_2 \geq 0, \\
\gamma_2 (x_2 (T) - \beta_T) = 0.
\end{cases}
\]

Without any further assumption on the production function we cannot continue the analysis, because the motion equation and the first adjoint equation are coupled. Therefore we focus on a special case, by adding the assumption that the production function is linear:

\[ f (x_1) = \lambda x_1, \quad \text{where } \lambda > 0. \]

As a consequence, the first adjoint equation is \( \dot{p}_1 (t) = -\lambda p_1 (t) \) and the adjoint system is uncoupled from the motion equation, so that the system results analytically manageable.

### 3.2 Limit instances \((p_0 = 0)\)

We first consider the possibility that \( p_0 = 0. \)

**Problem (3)**

If \( p_0 = 0 \) and \( \gamma_2 = 0 \) then \( \gamma_1 > 0 \) and therefore the transversality condition implies that \( x_1 (T) = \alpha_T. \) The solutions of the adjoint equations are \( p_1 (t) = \gamma_1 \exp (\lambda (T - t)) > 0 \) and \( p_2 (t) \equiv 0, \) hence the Hamiltonian function

\[ H (x_1, x_2, p_1, p_2, u, t) = p_1 (\lambda x_1 - u) \]

is maximized w.r.t. \( u \in [0, +\infty) \) by choosing \( u = 0. \) The resulting no consumption policy satisfies the transversality condition if and only if the associated capital stock level reaches the final value \( x_1 (T) = \alpha_T. \) This is true if and only if \( \alpha_T = \alpha_0 e^{\lambda T}. \)

**Problem (4)**

Symmetrically, if \( p_0 = 0 \) and \( \gamma_1 = 0 \) then \( \gamma_2 > 0, \) hence we get \( x_2 (T) = \beta_T. \) By the adjoint equations we have that \( p_2 (t) \equiv \gamma_2 > 0 \) and \( p_1 (t) \equiv 0, \) hence

\[ H (x_1, x_2, p_1, p_2, u, t) = p_2 e^{-\rho t} U (u) \]

which is upper unbounded w.r.t. \( u \in [0, +\infty). \) As the shadow price of the capital stock \( p_1 (t) \) is zero, unbounded consumption is suggested: there does not exist any optimal solution with \( p_0 = 0. \)
3.3 Degenerate instances

Now, let us assume that \( p_0 = 1 \); the transversality conditions for the problems (3) and (4) become the following, respectively:

\[
\begin{align*}
\begin{cases}
p_1 (T) = \gamma_1, \\
p_2 (T) = 1, \\
\gamma_1 \geq 0, \\
\gamma_1 (x_1 (T) - \alpha_T) = 0,
\end{cases}
\quad \begin{cases}
p_1 (T) = 1, \\
p_2 (T) = \gamma_2, \\
\gamma_2 \geq 0, \\
\gamma_2 (x_2 (T) - \beta_T) = 0.
\end{cases}
\end{align*}
\]

We consider first the trivial cases.

*Problem (3)*

If \( \gamma_1 = 0 \), then the adjoint equations give us \( p_1 (t) \equiv 0 \) and \( p_2 (t) \equiv 1 \); the Hamiltonian function becomes

\[
H(x_1, x_2, p_1, p_2, u, t) = e^{-\rho t} U(u),
\]

which is upper unbounded w.r.t. \( u \). This is the same situation as with \( p_0 = 0 \) for problem (4): there does not exist any optimal solution with \( p_0 = 1 \) and \( \gamma_1 = 0 \).

*Problem (4)*

Symmetrically we have: if \( \gamma_2 = 0 \), then the solution of the adjoint equations are \( p_1 (t) = \exp (\lambda (T - t)) > 0 \) and \( p_2 (t) \equiv 0 \); hence

\[
H(x_1, x_2, p_1, p_2, u, t) = p_1 (\lambda x_1 - u),
\]

which is maximized w.r.t. \( u \in [0, +\infty) \) by choosing \( u = 0 \). The resulting control \( u(t) \equiv 0 \) satisfies the final conditions if and only if \( \beta_T \leq 0 \), in which case only capital growth matters.

3.4 The controllability problem

Focusing on the nontrivial cases only, we arrive at the transversality conditions that can be written for the problems (3) and (4) respectively as

\[
\begin{align*}
\begin{cases}
p_1 (T) = \gamma_1 > 0, \\
p_2 (T) = 1, \\
x_1 (T) = \alpha_T,
\end{cases}
\quad \begin{cases}
p_1 (T) = 1, \\
p_2 (T) = \gamma_2 > 0, \\
x_2 (T) = \beta_T.
\end{cases}
\end{align*}
\]

These conditions, together with the adjoint equations, imply that \( p_1 (t) \) and \( p_2 (t) \) are strictly positive functions and their analytical forms for problem (3) and (4) respectively are:

\[
\begin{align*}
\begin{cases}
p_1 (t) = \gamma_1 e^{\lambda (T-t)} > 0, \\
p_2 (t) \equiv 1,
\end{cases}
\quad \begin{cases}
p_1 (t) = e^{\lambda (T-t)} > 0, \\
p_2 (t) \equiv \gamma_2 > 0.
\end{cases}
\end{align*}
\]
Now, we observe that the Hamiltonian function, which has to be maximized w.r.t. $u \in [0, +\infty)$, is a strictly concave function. Hence it has a unique maximum point, which is the solution of the equation

$$U'(u) = e^{\rho t} p_1(t) / p_2(t) .$$

(5)

After denoting by $\phi : (0, +\infty) \to (0, +\infty)$ the inverse function of $U'$, we obtain that the condition (5) implies that the optimal policy is

$$u^*(t) = \phi \left( \pi e^{(\rho - \lambda)t} \right),$$

where $\pi$ is a positive constant. Actually, the decision maker has to choose the constant $\pi \in (0, +\infty)$ in order to drive the system of ODEs

$$\begin{cases}
\dot{x}_1(t) = \lambda x_1(t) - u^*(t), \\
\dot{x}_2(t) = e^{-\rho t} U(u^*(t)),
\end{cases}$$

from the initial point

$$\begin{cases}
x_1(0) = \alpha_0, \\
x_2(0) = 0,
\end{cases}$$

to a point in either final surface $S_1 = \{\alpha_T\} \times \mathbb{R}$, for the standard problem (3), or $S_2 = \mathbb{R} \times \{\beta_T\}$, for the symmetric problem (4).

### 3.5 Pareto efficient solutions

Following the logical path of Section 2.5, let us define the functions $V_2(\alpha_T)$ and $V_1(\beta_T)$ as the optimal values of the objective functionals of the standard Ramsey’s problem, the former, which we call the immediate utility of final capital stock, and of the symmetric Ramsey’s problem, the latter, which we call the final capital stock level of immediate utility.

We note that the optimal control in both problems is characterized only by the value of the parameter $\pi$. Moreover, if we require that the system reaches an admissible final value for the state variable $x_1^*(T)$ ($x_2^*(T)$, respectively), then we find a unique $\pi$ which allows to satisfy the requirement. Such an observation allows to consider $x_1^*(T)$ and $x_2^*(T)$ as two 1-1 functions of $\pi$: hence the relation between $x_1^*(T)$ and $x_2^*(T)$ is a 1-1 function, which is $V_1$ in one direction and $V_2$ in the opposite one. Thus $V_2 = V_1^{-1}$. Their regularity follows from the utility function regularity. Finally,

- if $u^*(t)$ is an optimal control of the standard Ramsey’s problem, then it is optimal also for the symmetric Ramsey’s problem, provided that $\beta_T = V_2(\alpha_T)$;
- if $u^*(t)$ is an optimal control of the symmetric Ramsey’s problem, then it is optimal also for the standard Ramsey’s problem, provided that $\alpha_T = V_1(\beta_T)$.
4 Conclusion

In this paper we consider some symmetric problems in the framework of two classical economic models (i.e. consumption versus investment and Ramsey’s models). We prove that, from an analytical point of view, both the standard problem and the symmetric one can be reduced to the same controllability problem. The exploitation of the symmetry among problems is well known and studied in the mathematical optimization theory, but it is met less frequently in the economic analysis. Here we have shown that the solutions of two classical problems can be reinterpreted as the solutions of different problems which are symmetric to them and which have their own interesting economic meaning.

Acknowledgements

The authors wish to thank the anonymous Referees for requiring a crucial explanation and for suggesting the suitable quotation of Volpato, Giannessi and Morgantini.

References

On the coincidence of system optimum and user equilibrium for a widely used family of cost functions

Paolo Pellizzari and Annamaria Sorato

Dept. of Applied Mathematics
University "Ca' Foscari" of Venice
Dorsoduro 3825/E, 30123 Venice, Italy.

Abstract. In a simple two-node, one origin-destination network with multiple links, we characterize the coincidence of system optimum, that minimizes the total cost of agents with user equilibrium, that equalizes the cost in each (used) link. If cost functions are, up to a constant, homogeneous of the same degree then the system optimum and the user equilibrium are the same if and only if the freeflows are constant. Some examples show that the hypotheses are not redundant.

Keywords. Traffic flows, system optimum, user equilibrium.

JEL Classification: R41
MSC Classification: 90B20, 60K30

1 Introduction

We consider a model of traffic flow on a network whose roads are used by non-cooperative agents that travel from an origin to a destination, minimizing their own travel time or another suitable cost function. It is well know that the selfish user equilibrium in a traffic network can differ substantially from the globally optimal flow that makes the incurred total (or average) cost minimal. This is due to the different perspective used by the travellers: selfish agents select their route to minimize their own travel time with no regard to the congestion burden they put on others. As this externality is not taken into account, the resulting user equilibrium can be largely different from a system optimum, that is an efficient global allocation minimizing the total cost in the network. This feature is important for the clever design of new or improved networks, in the sense that the behaviour of the users must be taken into consideration, as the Braess paradox strikingly shows, [3] or chapter 6 in [1] for a colloquial presentation whose terminology is repeated in this paper. Along the same research line, other curious and puzzling effects are documented in [4] and [5] where it is shown that

* We thank Marco Li Calzi for drawing our attention on homogeneous cost functions and two anonymous referees for useful remarks. All errors are (sadly) ours: paolop@unive.it, amsorato@unive.it
polluting emissions can increase even when traffic demand decreases or when a link with zero emissions is added to certain networks.

Recent research work has compared the two equilibria, trying to quantify the extent of the increment of total travelling cost due to the selfish “lack of regulation”, as it is dubbed in [8]. Clearly, a system optimum can be obtained enforcing the users to select the appropriate paths, but this has huge coordination costs and would inevitably cause reactions by the agents that are routed along costly (or time consuming) roads. It is in fact very unlikely that an agent tolerates higher costs for the sake of social benefit, when other “equal” users have smaller costs. Enforcement seldom being a useful and realistic policy, efforts have aimed to suggest alternative paths to travellers by radio broadcasts or various signalling devices and adoption of pricing policies.

As theoretical and experimental work, [9], stress the ubiquitousness of sub-optimal selfish equilibria, we are interested in this paper in exploring situations in which the system optimum and the user equilibrium are the same. In detail, we investigate the properties a traffic network should possess to ensure that the two equilibria coincide. This is of obvious interest to a central planner that might modify the network (for example adding or deleting links and adjusting the link costs) in order to exploit the greedy behaviour of the users to achieve maximal efficiency with no visible action (other than proper network construction or modification) or enforcement. To the best of our knowledge, determination of conditions such that system optima are the same than user equilibria has received no attention in the literature.

We address the problem in a simplified setting with only two nodes, one origin-destination pair and multiple paths, that in this case can also be thought as edges or links. The situation is of interest in cases where a heavy traffic demand from the origin to the destination splits in various routes with no common intersection. The simple structure of the network indeed allows Selten and coauthors to develop an experimental study in an identical framework that is studied also in [7].

Our main results shows that from same degree homogeneity of the cost functions up to a constant, it follows that positive system optima are also user equilibria if and only if there are constant freeflows (the cost incurred when the flow on the link is null).

The rest of paper is organized as follows. Section 2 describes the model of a traffic problem on a network in a rather standard way, defines the user equilibrium (UE) and the system optimum (SO) and gives some useful lemmas. In the following Section, we state our main result for a two-nodes network with homogeneous costs up to a constant. Finally, Section 4 is devoted to provide some examples and remarks.

2 The model

We consider a directed network $G = (V, E)$ with vertex set $V$ and edge set $E$. We restrict our attention to the simple case of unique source-target pair,
setting \( V = \{O, T\} \) and assume there are \( k \) edges (roads) linking \( O \) to \( T \). The contribution of each single agent travelling on \( G \) is negligible and we assume that unitary traffic is flowing on the network. Define a flow \( x \) as a \( k \)-tuple of nonnegative real values \( x = (x_1, x_2, \ldots, x_k) \) such that \( \sum_{j=1}^{k} x_j = 1 \), where each \( x_j \) is the fraction of (total) load travelling along the \( j \)-th edge. This description is, up to a normalization, equivalent to the widely used approach that prescribes some total travel demand \( N \) in such a way that \( q_j \) travellers are on \( j \)-th path and \( \sum_{j} q_j = N \). Each edge \( e \in E = \{e_1, e_2, \ldots, e_k\} \) is equipped with a differentiable and nondecreasing cost function \( C_j(x) \). We assume that the form of the cost function on one edge takes the form \( C_j(x) = a_j + D_j(x) \), with \( D_j(0) = 0 \) and \( a_j \geq 0 \), in such a way that the cost can be split in two parts: the constant \( a_j \), often named freeflow, is the cost incurred when null flow is travelling on the arc, and the term \( D_j(x) \) penalizing heavy load. Trivially, the functions \( D_j \) inherit the smoothness properties of the costs \( C_j \) and are differentiable and nondecreasing.

Figure 1 shows a stylized graph of the network.

There are two interesting situations in such a network: a selfish user equilibrium, where agents equally spread congestion on all used links and an efficient system optimum that minimizes the sum (or, equivalently, the mean) of the incurred costs.

**Definition 1 (User Equilibrium).** A nonnegative flow \( \mathbf{x} = (x_1, \ldots, x_k) \) is a User Equilibrium (UE) if for all \( i, j \in \{1, \ldots, k\} \) such that \( \bar{x}_i > 0 \) and \( \delta \in [0, \bar{x}_i] \), we have \( \sum_j \bar{x}_j = 1 \) and

\[
C_i(x_i) \leq C_j(x_j + \delta).
\]
Loosely speaking, for each pair of edges, one’s cost increases moving from one edge to another. In particular, if $\bar{x}$ is a UE then all used edges have the same cost, [6].

**Lemma 1.** Let $\bar{x}$ be an UE. Then

$$C_j(\bar{x}_j) = \lambda,$$

for all $j$ such that $\bar{x}_j > 0$.

**Definition 2 (System Optimum).** A flow $\mathbf{x}^* = (x_1^*, \ldots, x_k^*)$ is a System Optimum (SO) if it solves the optimization problem

$$\min_{x_1, \ldots, x_k} \mathbf{C}(\mathbf{x}) = \sum_{j=1}^k x_j C_j(x_j).$$

The existence of a UE was established in [2], assuming that the cost function of each network link is continuous and nondecreasing in the flow. Moreover, boundedness of the feasible region and continuity of cost functions ensure, by Weierstrass theorem, that a SO exists.

In the next section we provide sufficient and necessary conditions for an UE to be also a SO and viceversa. The following lemma will be useful in the sequel.

**Lemma 2.** Let the cost functions be such that, for every $j = 1, \ldots, k$

$$C_j(x) > a_j, \text{ for all } x > 0;$$

$$C_j(0) = a, \forall j.$$

Then

A1. The UE $\mathbf{x}$ is such that $\bar{x}_j > 0, j = 1, \ldots, k$, i.e. all links are used by the agents. In brief, $\mathbf{x} > 0$.

A2. The SO $\mathbf{x}^*$ is such that $x^*_j > 0, j = 1, \ldots, k$, i.e. nonnegative constraints are not binding. In brief, $\mathbf{x}^* > 0$.

**Proof.** Let $\mathbf{x} = (\bar{x}_1, \ldots, \bar{x}_k)$ be an UE. By contradiction, we assume without loss of generality that $\bar{x}_1 = 0$ and $\bar{x}_2 > 0$. Noting that $D_2(\bar{x}_2) > 0$, by continuity of $D_1$, there exists $\delta$ such that

$$\forall x \in [0, \delta], \quad 0 < D_1(x) < D_2(\bar{x}_2).$$

Hence, there exists a positive $\delta < \bar{x}_2$ such that

$$C_2(\bar{x}_2) > C_1(\bar{x}_1 + \delta) = C_1(\delta),$$

contradicting the assumption that $\mathbf{x}$ is UE.
Let us now assume that \( x^* = (x_1^*, \ldots, x_k^*) \) be a SO. By contradiction, we assume without loss of generality that \( x_1^* = 0 \) and \( x_2^* > 0 \). We will show that \( x^* \) cannot be optimal showing that moving some amount \( \epsilon \) of flow from link 2 to 1 reduces the objective function. In other words, we claim that

\[
C(\epsilon, x_2^* - \epsilon, x_3^*, \ldots, x_k^*) < C(x^*) \quad \text{for some} \quad \epsilon > 0.
\]

Consider the sum of the first two terms in the summation (1)

\[
x_1^* C_1(x_1^*) + x_2^* C_2(x_2^*) = x_2^*(a + D_2(x_2^*))
\]

to be compared to

\[
\epsilon C_1(\epsilon) + (x_2^* - \epsilon) C_2(x_2^* - \epsilon) = \epsilon(a + D_1(\epsilon)) + (x_2^* - \epsilon)(a + D_2(x_2^* - \epsilon)).
\]

Using the definition of the \( C_j \)'s and a Taylor expansion, the right hand side yields

\[
\begin{align*}
&\quad \epsilon a + \epsilon (D_1(0) + \epsilon D_1'(0) + o(\epsilon)) + \\
&+ x_2^* a + x_2^* (D_2(x_2^*) - \epsilon D_2'(x_2^*) + o(\epsilon)) + \\
&- \epsilon a - \epsilon (D_2(x_2^*) - \epsilon D_2'(x_2^*) + o(\epsilon)).
\end{align*}
\]

Simplifying and omitting higher order terms we obtain

\[
x_2^*(a + D_2(x_2^*)) - \epsilon \left( x_2^* D_2'(x_2^*) + D_2(x_2^*) \right) > x_2^*(a + D_2(x_2^*)),
\]

contradicting the optimality of \( x^* \).

Some remarks are in order. A1 rules out the existence of links that are not used by selfish agents. This does not mean that the links with null flow in an UE are not used in a SO. Indeed, it is well known that there are situations where A2 holds even though A1 does not, as some social benefit can often be obtained rerouting a portion of the traffic on links that would not have been used on a selfish basis.

The previous result basically shows that if the \( a_j \)'s are constant then both \( \mathbf{x} \) and \( x^* \) have strictly positive components (i.e. A1 and A2 always hold). In the following section we give our main result, stating that in the presence of homogeneous \( D_j \)'s, positive SO and UE coincide if and only if the freeflows are constant.

## 3 Identity of SO and UE

This section is devoted to characterize the networks that admit coincident SO and UE's.
Proposition 1. Assume the functions $D_j$ be homogeneous of degree $p > 0$ on $\mathbb{R}^+$, nondecreasing and not identically null. Then $x > 0$ is a SO and a UE if and only if $a_j \equiv a, j = 1, \ldots, k$ i.e. the $a_j$’s are constant.

Proof. The assumptions of homogeneous, nondecreasing and not identically null $D_j$’s imply that (see [10])

$$D_j(x) = \beta_j x^p, \beta_j > 0, j = 1, \ldots, k.$$ 

In particular, as each term $x_j C_j$ is a convex function, it follows that the objective function $C(x)$ in (1) is convex. Hence, being convex the optimization problem for the SO, the Kuhn-Tucker (KT) first order conditions are necessary and sufficient for a solution, see [11].

Assume $x > 0$ is both a SO and an UE. Then by Lemma 1 we have

$$a_j + D_j(x_j) = \lambda, j = 1, \ldots, k \sum_{j=1}^k x_j = 1. \quad (2)$$

As $x$ is SO also KT conditions must hold, hence

$$a_j + D_j(x_j) + x_j D_j'(x_j) - \xi - \mu_j = 0, \quad (3)$$

where $\xi, \mu_1, \ldots, \mu_k$ are suitable multipliers. Using Euler’s theorem for homogeneous functions and observing that $x > 0$ implies $\mu_j = 0, j = 1, \ldots, k$, equation (3) can be rewritten as

$$a_j + (1 + p)D_j(x_j) - \xi = 0.$$ 

Recalling from (2) that $D_j(x_j) = \lambda - a_j$ and substituting in the previous equation gives

$$a_j + (1 + p)(\lambda - a_j) - \xi = 0, \quad j = 1, \ldots, k.$$ 

Hence

$$a_j = \frac{\lambda(1 + p) - \xi}{p},$$

which is a constant independent of $j$.

Assume now that $a_j \equiv a$ and let $x^*$ be a SO: we want to show that it is also a UE and $x^* > 0$. As before, the KT conditions for the problem (1) hold at $x^*$:

$$a + (1 + p)D_j(x^*_j) - \xi - \mu_j = 0, \quad (4)$$
By Lemma 2, $x^* > 0$ and this in turn gives $\mu_j = 0$, $j = 1, \ldots, k$. Then $D_j(x^*_j) = (\xi - a)/(1 + p)$ for all $j$ and $x^*$ is also an UE by Lemma 1.

Finally, assume that $\mathbf{x}$ is an UE (and still $a_j \equiv a$). Strict positiveness of $\mathbf{x}$ is an immediate consequence of Lemma 2. Then $D_j(\pi_j) = \lambda$ for all $j$. We want to show that,

$$(\mathbf{x}, \xi, \mu_1, \ldots, \mu_k) = (\pi_1, \ldots, \pi_k, \xi, 0, \ldots, 0)$$

for some $\xi$, (5) is a full solution of KT conditions for a SO. Plugging (5) into the KT conditions gives

$$a + (1 + p)D_j(\pi_j) - \xi = 0,$$

$$\sum_{j=1}^k \pi_j = 1,$$

$$\pi_j \geq 0, \mu_j \geq 0,$$

$$\pi_j \mu_j = 0,$$

$$\forall j = 1, \ldots, k,$$

that is trivially satisfied if we set $\xi = (1 + p)\lambda + a$. By sufficiency of KT conditions, we can conclude that $\mathbf{x}$ is a SO. \hfill \Box

4 Examples and discussion

This section provides some simple (counter)examples to demonstrate that all the assumptions are required and gives some conclusive remarks.

**Example 1.** Let $C_1(x) = 10 + 10x$, $C_2(x) = 30 + 10x$. Then it is immediate to show that the UE and SO are identical, $\mathbf{x} = (1, 0) = x^*$. Though UE coincides with SO, $x^*$ is not positive and Proposition 1 does not hold being $a_1 = 10 \neq 30 = a_2$.

It is very easy to find networks with non constant freeflows that have different SO and UE, but even in the (rare) cases when $x^* = \mathbf{x}$, Example 1 shows that $x^*$ is not necessarily positive.

The following example stresses that the assumption of homogeneity of same degree $p$ is essential.

**Example 2.** Let $C_1(x) = 10 + 10x$, $C_2(x) = 25/2 + 10x^2$. Some computations show that

$$\mathbf{x} = \left(\frac{1}{2}, \frac{1}{2}\right) = x^* > 0,$$
but nevertheless the $a_j$’s are different.

We conclude the paper with a final consideration. Homogeneity of the $D_j$’s is a rather acceptable and widely used assumption on the cost functions. With some additional technical hypotheses, we characterize the identity of SO and UE completely in terms of constant freeflows. Most of the traffic networks have however widely different freeflows and this is in total agreement with the well known fact that coincidence of UE and SO is a very rare event. Proposition 1 could hopefully suggest proper behaviour when links of an existing network are modified or added by a central planner. If efficiency is desired, then an effort should be done in order to achieve constant freeflows as this situation would ‘force’ unaware selfish agents to adopt socially optimal flows without any explicit form of imposition.

References