View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

centrA:
Fundación Centro de Estudios Andaluces

Documento de Trabajo
Serie Economía E2003/08

The Impossibility of Strategy-Proof Clustering[*]

Javier Perote Peña
Universidad Rey Juan Carlos

Juan Perote Peña
Universidad Pablo de Olavide de Sevilla

RESUMEN

Los métodos de ''clustering'' agrupan individuos u objetos de acuerdo con la información de que se dispone acerca de su parecido o proximidad. Cuando la información básica para generar los grupos no puede ser fácilmente observada o verificada, el diseñador de los grupos debe apoyarse en información procedente de los individuos que se encuentran detrás de las observaciones. Cuando los individuos reciben utilidad de una decisión pública tomada con datos agregados dentro del grupo de cada uno y además tienen preferencias unimodales, probamos que no existen métodos de "clustering" tales que el comportamiento de revelación de la verdad es siempre una estrategia dominante.

Palabras clave: métodos de agrupamiento, a prueba de estrategias, preferencias unimodales, decisión pública.

ABSTRACT

Clustering methods group individuals or objects based on information about their similarity or proximity. When the raw information to generate the clusters cannot be easily observed or verified, the clusters designer must rely on information reported on individuals behind the observations. When individuals receive utility from a public decision taken with aggregated data within each own's cluster and have single-peaked preferences, we prove that there do not exist cluster methods such that truth-revealing behavior is always a dominant strategy.

Keywords: clustering methods, strategy-proofness, single-peaked preferences, public decision.
JEL classification: C44 y D70.

[*] The authors wish to acknowledge to the Fundación Centro de Estudios Andaluces (centrA) for financial support.

The Impossibility of Strategy-proof Clustering[*]

Juan Perote-Peña and Javier Perote

Authors: Juan Perote-Peña and Javier Perote.

Author 1: Juan Perote-Peña

A*ffiliation:* Departamento de Economia y Empresa, Universidad Pablo de Olavide de Sevilla y Fundación Centro de Estudios Andaluces (CentrA, internal researcher in the Microeconomics group).

Running title: Strategy-proof clustering.

Address for manuscript correspondence:

Juan Perote Peña, Departamento de Economía y Empresa, Universidad Pablo de Olavide de Sevilla, Carretera de Utrera, Km. 1, 41013, Sevilla, Spain.

Telephone: +34-954349187.

E-mail: jperpen@dee.upo.es

Fax: +34-954349339

Author 2: Javier Perote

A*ffiliation:* Facultad de C.C. Jurídicas y Sociales, Universidad Rey Juan Carlos.

Address: Universidad Rey Juan Carlos, Facultad de C.C. Jurídicas y Sociales, Campus de Vicálvaro, P^{o}de los Astilleros s/n, 28032 Madrid, Spain.

Telephone: +34-913019923.

E-mail: perote@fcjs.urjc.es

[*]We wish to acknowledge to the Fundación Centro de Estudios Andaluces (CentrA) for financial support.

Abstract

Clustering methods group individuals or objects based on information about their similarity or proximity. When the raw information to generate the clusters cannot be easily observed or verified, the clusters designer must rely on information reported on individuals behind the observations. When individuals receive utility from a public decision taken with aggregated data within each own's cluster and have single-peaked preferences, we prove that there do not exist cluster methods such that truth-revealing behavior is always a dominant strategy.

JEL classification numbers: C44, D70.

Keywords: clustering methods, strategy-proofness, single-peaked preferences, public decision.

Resumen

Los métodos de "*clustering*" agrupan individuos u objetos en base a información acerca de su parecido o proximidad. Cuando la información básica para generar los grupos no puede ser fácilmente observada o verificada, el diseñador de los grupos debe apoyarse en información procedente de los individuos que se encuentran detrás de las observaciones. Cuando los individuos reciben utilidad de una decisión pública tomada con datos agregados dentro del grupo de cada uno y además tienen preferencias unimodales, probamos que no existen métodos de *clustering* tales que el comportamiento de revelación de la verdad es siempre una estrategia dominante.

1 Introduction

Let us imagine a society in which the agents or individuals live in some locations on a given region. A social planner must efficiently locate some public facilities -say, health services, police stations, schools, etc.- in the region based on reported information about each individual location. The individuals would like the public facility to be located on their own location -to minimize the transport costs, for example- and have preferences on the set of feasible locations that are continuous and single-peaked: the further the location is along a straight line starting from the single most preferred alternative -top location or peak, i.e., the agent's own location-, the worse it is for the agent. When individuals are asked to report their locations, they might have an incentive to lie and report false information in order to achieve a better location for themselves. When just one public facility must be chosen, the allocation problem can be solved in some cases. For example, in the one-dimensional case it is possible to design Pareto-optimal allocation methods such that no individual will ever have an incentive to report false information about her own location -see Moulin (1980)-. In the real world usually more than just one public facility can be located and in this case, there are different possibilities open. For example, individuals might be allowed to choose the service point they like the most from a fixed amount of points to be chosen by society -see, for example, Miyagawa (1998, 2001)-. Other possibility consists in dividing society into different areas that group the set of individuals associated to each service point. In this note, we assume that the social planner has no clear constraint in the number of public facilities to locate, but would like to group individuals that live "close" to each other to allocate service points with a minimum of transport costs. Since the only information the planner has to build the groups or "neighborhoods" is about reported locations, she applies a clustering technique to those points generating the different areas that will deserve a service point of their own. For the problem to be meaningful we shall need to impose a minimal constraint on the admissible clustering techniques to avoid trivial results. Therefore, the extreme cases of always considering a unique cluster or as many clusters as individuals cannot be proper cluster techniques. After excluding these two extreme and trivial cases, we find that every other admissible clustering technique cannot rule out the possibility of individuals lying about their locations.

Notice that our model is not restricted to a spatial allocation problem, but admits a much wider range of economic problems. The information to be reported can be any measures of different socioeconomic variables of agents to be taken into account to create clusters. The next section is devoted to both the model and the main impossibility result.

2 The model and the result

Let $N = \{1, ..., n\}$ be a *society* composed by a finite number $n \geq 3$ of *agents* or *individuals* denoted by $i, j, h \in N$. Let π, π' denote different *partitions* of society, i.e., π is a reflexive, symmetric and transitive binary relation $\pi \in N \times N$. Let Π be the set of all possible partitions of society N. Therefore, $(i, j) \in \pi$ means that agent $j \in N$ belongs to the same group of agent $i \in N$. Notice that given any partition π, if we denote as $\pi_i = \{(i, j) \in \pi \mid j \in N\}\ \forall i \in N$, it must hold that $\forall i, j \in N$, either $\pi_i \cap \pi_j = \emptyset$ or $\pi_i = \pi_j$ and $\pi_i \neq \emptyset\ \forall i \in N$. Let E be the real line and let $x, y, z \in E^2$ denote arbitrary elements -locations- in E^2. Let $d(x, y) \in E_+$ denote the euclidean *distance* between the points or locations x and y. Each individual $i \in N$ has a *preference relation* defined on the set E^2 denoted as R_i, where P_i and I_i are the asymmetric and symmetric parts of R_i respectively. Some preference relations can be represented by means of utility functions. A *utility function* u is a function of the kind: $u : E^2 \rightarrow E$. A preference relation R_i is represented by a utility function u_i if $\forall x, y \in E^2,\ xR_iy \longleftrightarrow u_i(x) \geq u_i(y)$. We require admissible preferences to be continuous in E^2 and such that (1) and (2) hold.

(1). For every individual $i \in N$,there exists a unique location $p(R_i) \in E^2$ such that $p(R_i)P_ix\ \forall x \in E^2 \backslash p(R_i)$.

(2). For every individual $i \in N, \forall x \in E^2 \backslash p(R_i),\ \forall \lambda \in (0, 1)$,

$$[\lambda p(R_i) + (1 - \lambda)x]\, P_ix.$$

Let $\Re$ be the set of all continuous preferences on E^2 such that (1) and (2) hold[1]. An *economy* $\mathbf{R} \in \Re^n$ is a list of one admissible preference relation for every agent, and for any economy $\mathbf{R} \in \Re^n$ there will be a vector of associated peaks $p(\mathbf{R}) = (p(R_1), ..., p(R_n)) \in E^{2n}$.

A *clustering technique* is a function $C : E^{2n} \rightarrow \Pi$. A clustering technique is intended to put together as members of the same group or cluster the agents that are "similar" among themselves and separate them from those that are less similar, the definition of the similarity degree being implicit in function C. Therefore, $\forall \mathbf{x} = (x, y, ..., z) \in E^{2n}$, $C(\mathbf{x})$ is a possible partition of the set N, so whenever $(i, j) \in C(\mathbf{x})$, we say that agents $i \in N$ and $j \in N$ belong to the same cluster. Notice that the elements in E^{2n} used by function C to generate clusters are the reported "peaks" of the agents.

[1] Conditions (1) and (2) together with the assumption of continuity of preferences is a possible definition of the concept of "single-peakedness" when extended to two dimensions.

There are hundreds of clustering techniques proposed in the literature -see Everitt (1993) for an excellent survey, for example-, each one characterized by different concepts of the "distance" between groups of individuals and the way they are formed, either using algorithms or optimization techniques. We only need a very loose characterization of admissible clustering techniques that is fulfilled by most of them.

Definition 1 *A clustering technique C is **admissible** if the following two conditions hold:*

(i). $\exists \widehat{d} \in E_{++}$*, such that* $\forall \mathbf{x} = (x_i, x_j, x_{-i-j}) \in E^{2n}$ *such that* $\exists i \in N$ *with* $d(x_i, x_j) \geq \widehat{d}\, \forall j \in N \rightarrow (j,i) \notin C(\mathbf{x})\, \forall j \in N$.

(ii). $\exists \widetilde{d} \in E_{++}$*, such that* $\forall \mathbf{x} = (x_i, x_j, x_{-i-j}) \in E^{2n}$ *such that* $\exists i, j \in N$ *with* $d(x_i, x_j) \leq \widetilde{d} \rightarrow (j,i) \in C(\mathbf{x})$.

Condition (i) implies that when an agent's peak is far enough from any other's peak, it deserves to form an independent cluster. Analogously, condition (ii) means that when two agents' peaks are close enough they must belong to the same cluster. Notice that, in particular, (ii) guarantees that any two agents with preferences such that their peaks are located in the same place must always be in the same cluster. Now, we provide a very general model for the public choice taken in each cluster. We only assume that the choice must be efficient when considering the members of the cluster alone -they might vote or reach agreements to locate the public service.

Definition 2 *Given any clustering technique C, a **public location function** is a function* $L : \Re^n \rightarrow E^{2n}$ *such that* $\forall \mathbf{R} \in \Re^n$*,with reported*[2] *vector of peaks* $\mathbf{x} = (x_1, x_2, ..., x_n) \in E^{2n}$, $L(\mathbf{R}) = (L_1(\mathbf{R}), ..., L_n(\mathbf{R})) \in E^{2n}$ *is such that the following two conditions hold:*

(i'). $\forall i, j \in N, (i,j) \in C(\mathbf{x}) \rightarrow L_i(\mathbf{R}) = L_j(\mathbf{R})$.

(ii'). $\forall i \in N$*, there does not exist* $z \in E^2$ *such that* $z P_j L_i(\mathbf{R})\, \forall j \in N$ *such that* $(i,j) \in C(\mathbf{x})$.

A public location function establishes a common Pareto-efficient location for the public service for all members within each cluster.

Definition 3 *An **allocation method** in our model is a pair (C, L) where C is an admissible clustering technique and L is a public location function.*

Now, we define the strategic property we are interested in:

[2] Notice that the declared or reported peaks of the agents do not need to coincide with the true peaks, but they are the only information that the planner has to make the public location decision.

Definition 4 *An allocation method (C, L) is **manipulable** at economy $\mathbf{R} = (R_i, \mathbf{R}_{-i}) \in \Re^n$ with reported vector of peaks $\mathbf{x} = (x_1, x_2, ..., x_n) \in E^{2n}$, by individual $i \in N$, if there exists a preference relation $R_i' \in \Re$ with peak $x_i' \in E^2$ such that $L_i(R_i', \mathbf{R}_{-i}) P_i L_i(\mathbf{R})$.*

Definition 5 *An allocation method (C, L) is **strategy-proof** if it is not manipulable at any economy $\mathbf{R} \in \Re^n$ by any individual $i \in N$.*

We are now prepared to state our result.

Theorem 6 *There are no strategy-proof allocation methods under our assumptions*

Proof. We prove it by contradiction. Let (C, L) be a strategy-proof allocation method. Let us consider any two individuals $i, j \in N$ and the following economy $\mathbf{R} = (R_i, R_j, \mathbf{R}_{-i-j}) \in \Re^n$: $R_i \in \Re$, with peak $p(R_i)$ at $(0, 0)$ is any preference relation that can be represented by the following utility function: $\forall x = (x^1, x^2) \in E^2$,

$$u_i(x^1, x^2) = \begin{cases} -sign(x^2)x^2 + bx^1 & \text{if } x^1 \leq 0 \\ -sign(x^2)x^2 - cx^1 & \text{if } x^1 \geq 0 \end{cases} \tag{1}$$

where $b = \dfrac{\widetilde{\delta}}{2\left(\widehat{\delta} + \varepsilon\right)}$ and $c = 1$ ($\varepsilon > 0$ is any positive number). It is easy to check that preferences $R_i \in \Re$ above have indifference curves on the plane that correspond to a rhomboid with center in agent $i's$ peak - the origin $(0, 0)$- like the one depicted in *Figure 1.* The axis of the rhombus coincide with both x^1 and x^2. When we have $b = c = 1$, we obtain a symmetric rhombus. Notice that parameter b determines the slopes of the rhomboid left sides (for negative values of x^1) and parameter c determines the slopes of the rhomboid for positive values of x^2. The number $-a$ (the intercept of the rhomboid for negative values of x^2) is used as the measure of the utility that generates utility function $u_i(x^1, x^2)$, which is in fact a pyramid generating the rhomboids as level curves.

[Insert *Figure 1* about here]

Agent $j's$ preferences $R_j \in \Re$, with peak $p(R_j)$ at $(\widetilde{d}, 0)$ is any preference relation that can be represented by the following utility function: $\forall x = (x^1, x^2) \in E^2$,

$$u_j(x^1, x^2) = \begin{cases} -sign(x^2)x^2 + b(x^1 - \widetilde{d}) & \text{if } x^1 \leq \widetilde{d} \\ -sign(x^2)x^2 - c(x^1 - \widetilde{d}) & \text{if } x^1 \geq \widetilde{d} \end{cases} \tag{2}$$

where $b = 1$ and $c = \dfrac{\widetilde{\delta}}{2\left(\widehat{\delta}+\varepsilon\right)}$ ($\varepsilon > 0$ is any positive number). This preference relation is analogous to the former one, the indifferent curves being rhomboids centered at $(x^1, x^2) = (\widetilde{d}, 0)$. The remaining agents $h \in N \backslash \{i, j\}$ have the same preferences $R_h \in \Re$, with peak $p(R_h)$ at $(\widetilde{d} + \widehat{d}, 0)$, that can be represented by the following utility function: $\forall x = (x^1, x^2) \in E^2$,

$$u_h(x^1, x^2) = \left\{ \begin{array}{ll} -sign(x^2)x^2 + b(x^1 - \left[\widetilde{d}+\widehat{d}\right]) & \text{if } x^1 \leq \widetilde{d}+\widehat{d} \\ -sign(x^2)x^2 - c(x^1 - \left[\widetilde{d}+\widehat{d}\right]) & \text{if } x^1 \geq \widetilde{d}+\widehat{d} \end{array} \right. \tag{3}$$

Again, with $b = c = 1$. Let $\mathbf{x} \in E^{2n}$ be the vector of peaks of economy $\mathbf{R} \in \Re^n$, i.e., $\mathbf{x} = (x_i, x_j, \mathbf{x}_{-i-j})$. By construction, (i) and (ii), it holds for $\mathbf{R} = (R_i, R_j, \mathbf{R}_{-i-j}) \in \Re^n$ that $(i, j) \in C(\mathbf{x})$ and $\forall h, k \in N \backslash \{i, j\}$, it holds that $(h, k) \in C(\mathbf{x})$ by (ii), while $(i, h) \notin C(\mathbf{x})$ $\forall h \in N \backslash \{i, j\}$, so since $C(\mathbf{x})$ is a partition of N, it must be the case of $(j, h) \notin C(\mathbf{x})$ $\forall h \in N \backslash \{i, j\}$. Now, we prove that L must be such that $L_i(\mathbf{R}) = L_j(\mathbf{R}) \in \left[(0, 0), (\widetilde{d}, 0)\right]$ by (i') and (ii'). Given preferences $R_i, R_j \in \Re$, the only Pareto-optimal locations are on the straight line connecting both peaks -a segment lying on the x axis-: any other location $(x^1, x^2) \in E^2$ such that $x^1 < 0$ or $x^1 > \widetilde{d}$ are Pareto-dominated when considering only members of the cluster $i, j \in N$ by $(x^1, x^2) = (0, 0)$ and $(x^1, x^2) = (\widetilde{d}, 0)$ respectively and any other location such that $0 \leq x^1 \leq \widetilde{d}$ and $x^2 \neq 0$ is dominated by location $(x^1, 0)$ -both agents i and j attain a higher utility level given their preferences in (1) and (2) at the beginning of the proof-. Now, there are only two possibilities to consider: **Case 1**: $L_i(\mathbf{R}) = L_j(\mathbf{R}) = (\overline{x}^1, 0)$ with $\dfrac{\widetilde{d}}{2} \leq \overline{x}^1 \leq \widetilde{d}$. Then, we can imagine the admissible economy $\mathbf{R}' = (R_i', \mathbf{R}_{-i}) \in \Re^n$ where all preferences are the same as those in $\mathbf{R} \in \Re$ with the exception of individual $i's$ new preferences $R_i' \in \Re$, which can be represented by the following utility function u_i': $\forall x = (x^1, x^2) \in E^2$,

$$u_i'(x^1, x^2) = \left\{ \begin{array}{ll} -sign(x^2)x^2 + b(x^1 + \widehat{\delta}) & \text{if } x^1 \leq 0 \\ -sign(x^2)x^2 - c(x^1 + \widehat{\delta}) & \text{if } x^1 \geq 0 \end{array} \right. \tag{4}$$

where $c = b = 1$. This preference relation is a rhomboid centered at $(-\widehat{\delta}, 0)$, so $p(R_i') = (-\widehat{\delta}, 0)$. First of all, notice that given preferences $R_i \in \Re$, it holds that $u_i(-\widehat{\delta}, 0) > u_i(\overline{x}^1, 0)$, or, in other words,

$$(-\widehat{\delta}, 0) P_i (\overline{x}^1, 0) \tag{5}$$

Now, let us consider economy $\mathbf{R}' = (R_i', \mathbf{R}_{-i}) \in \Re^n$. Let $\mathbf{x}' \in E^{2n}$ be the vector of peaks associated to economy $\mathbf{R}' \in \Re^n$. Therefore, $x_i' = p(R_i') = (-\widehat{\delta}, 0)$, $x_j' = p(R_j) = (\widetilde{\delta}, 0)$ and $x_h' = p(R_h) = (\widetilde{\delta} + \widehat{\delta}, 0)$ $\forall h \in N \backslash \{i, j\}$. Now, notice that by (i), $(i, j), (i, h) \notin C(\mathbf{x}')$ $\forall h \in N \backslash \{i, j\}$, so agent $i \in N$ form her own cluster in economy $\mathbf{R}' \in \Re^n$ and by (i') and (ii') above, it must be the case of $L_i(\mathbf{R}') = p(R_i') = (-\widehat{\delta}, 0)$, while the public location for agent $i \in N$ associated to economy $\mathbf{R} \in \Re^n$ is by assumption $L_i(\mathbf{R}) = (\overline{x}^1, 0)$. Now, (5) above can be written as: $L_i(\mathbf{R}') = L_i(R_i', \mathbf{R}_{-i}) P_i L_i(\mathbf{R}')$, so the allocation method (C, L) is manipulable at economy $\mathbf{R} \in \Re^n$ with vector of peaks $\mathbf{x} \in E^{2n}$ by individual $i \in N$ by means of reporting false preferences $R_i' \in \Re$. Hence, (C, L) is not strategy-proof, entering into contradiction with our initial assumption. The only case left is **Case 2**: $L_i(\mathbf{R}) = L_j(\mathbf{R}) = (\overline{x}^1, 0)$ with $\frac{\widetilde{d}}{2} \geq \overline{x}^1 \geq 0$. If this is the case, let us consider the following economy: $\mathbf{R}'' = (R_j', \mathbf{R}_{-j}) \in \Re^n$, with associated vector of peaks $\mathbf{x}'' \in E^{2n}$ where all agents with the exception of $j \in N$ have the same preferences they had in economy $\mathbf{R} \in \Re^n$, and agent j's new preferences are: $R_j' = R_h$ $\forall h \in N \backslash \{i, j\}$ (i.e., those represented by utility function (3) above). The vector of peaks $\mathbf{x}'' \in E^{2n}$ will then be: $\mathbf{x}'' = (x_i, x_j', \mathbf{x}_{-i-j}) = ((0, 0), (\widetilde{\delta} + \widehat{\delta}, 0), ..., (\widetilde{\delta} + \widehat{\delta}, 0))$. By (i) and (ii), it holds for economy $\mathbf{R}'' \in \Re^n$ that $(i, j), (i, h) \notin C(\mathbf{x}'')$ $\forall h \in N \backslash \{i, j\}$ and $\forall h, k \neq i$, $(h, k) \in C(\mathbf{x}'')$. By (i') and (ii'), $L_j(\mathbf{R}'') = p(R_j') = p(R_h) = (\widetilde{\delta} + \widehat{\delta}, 0)$. Now, notice that given preferences $R_j \in \Re$, it holds that $u_j(\widetilde{\delta} + \widehat{\delta}, 0) > u_i(\overline{x}^1, 0)$, -see *Figure 2*- or, in other words,

$$(\widetilde{\delta} + \widehat{\delta}, 0) P_i (\overline{x}^1, 0) \tag{6}$$

Now, notice that expression (6) above can be written as: $L_j(\mathbf{R}'') = L_j(R_j', \mathbf{R}_{-j}) P_i L_j(\mathbf{R})$, so allocation method (C, L) is manipulable at economy $\mathbf{R} \in \Re^n$ with vector of peaks $\mathbf{x} \in E^{2n}$ by individual $j \in N$ by means of reporting false preferences $R_j' \in \Re$. Thus, (C, L) is not strategy-proof in this case either, so we cannot avoid contradictions and the theorem is proved. ■

[Insert *Figure 2* about here]

References

[1] Barberà, S., & M. Jackson (1994): "A characterization of Strategy- Proof Social Choice Functions for Economies with Pure Public Goods". Social Choice and Welfare 11, 241-252.

[2] Brams, S.J., Jones, M.A. & Kilgour, D.M. (2001): "Single-Peakedness and Disconnected Coalitions". Journal of Theoretical Politics, forthcoming.

[3] Everitt, B.S.: *Cluster Analysis*, London: Arnold cop., 1993.

[4] Miyagawa, E (1998): "Strategy-proof provision of a menu". Mimeo, Columbia University.

[5] Miyagawa, E (2001): "Locating libraries on a street". Social Choice and Welfare 18, 527-541.

[6] Moulin, H. (1980): "On Strategy- proofness and Single- peakedness". Public Choice 35, 437-455.

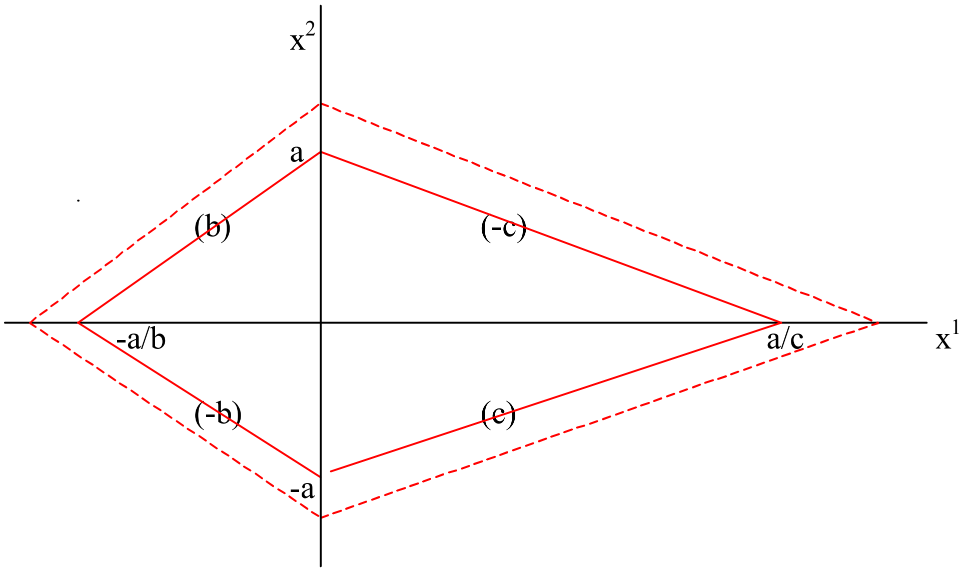

Figure 1:

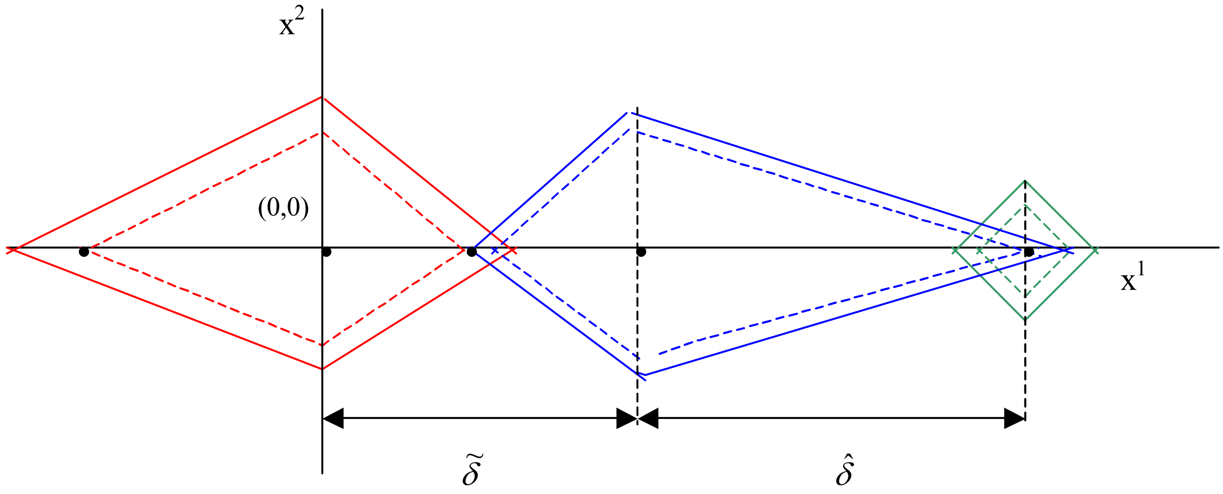

Figure 2: