Faculteit der Economische Wetenschappen en Econometrie

Serie Research Memoranda

On the Estimation of Stochastic Linear Relations*

B. Hanzon

Research Memorandum 1991-30
April 1991

vrije Universiteit *amsterdam*

On the estimation of stochastic linear relations*

Bernard Hanzon†
Dept. Econometrics, Free University Amsterdam

March 1991

Abstract

It is argued by various authors that the usual splitting of the set of variables in endogenous variables and exogenous variables (including instruments), should *not* be done *a priori*, i.e. before estimation, but *a posteriori*. The idea is that the estimation procedure should produce the possible splittings in endogenous and exogenous variables.

In this paper we try to make a first step in this direction by considering the (linear) *relations* between the variables as being stochastic, instead of the variables themselves. One then has the freedom to fix a number of variables (the exogenous variables, including the instruments) and as a result the remaining variables (the endogenous variables) are stochastic.

This is worked out by making the sets of linear relations, the so-called Grassmannians, into a metric space and using a generalization of Gaussian densities to such spaces. An important technical aspect of the analysis is the representation of the elements of a Grassmannian by symmetric idempotent matrices, also called orthogonal projection matrices. In the case of two variables it is shown that the densities that we use are in fact the Von Mises densities. For that case the Maximum Likelihood Estimators are derived. Remarks are made about the M.L.E. in the general case.

*Paper submitted to the ESEM91, Cambridge, U.K.

†Address: De Boelelaan 1105, 1081 HV Amsterdam, Holland; E-mail: bhnz@sara.nl

1 Introduction

In the standard multiple regression model one distinguishes a priori (stochastic) endogenous variables and (deterministic or at least "predetermined") exogenous variables, including instruments. In many econometric models the choice of what is endogenous and what is exogenous is rather arbitrary. For one thing many of the exogenous variables are only taken to be exogenous because it is decided that for the application in mind it is not necessary or appropriate to model this variable. If at some other stage of the modelling process it is decided to model such a variable after all one speaks of "endogenization" of the variable! Another class of exogenous variables is formed by the socalled instruments, or control variables. At first it seems that one could argue that such variables are truly exogenous because their value is determined by some decision maker and its value is not determined by the market etc. However this argument clearly depends very much on *who* analyses the economic phenomena under consideration, because if this is not the decision maker, the argument becomes a weak one, as in economics most variables are in the end the result of human decisions. Furthermore there is a more fundamental issue: by declaring certain variables to be instruments or exogenous one implicitly states a proposition about the model and this proposition is usually not verified empirically. The proposition is that the variable involved can be chosen freely, i. e. can take any value in its domain and is not restricted by the model equations. This point has been taken up by Willems [6,7,8] in a deterministic context.

To avoid the a priori choice of what are endogenous variables and what are exogenous variables and instruments, several approaches can be taken. One of them is to assume the exogenous variables are stochastic and to treat them in the same way as the endogenous variables are treated as far as the estimation procedure is concerned. This leads to "errors-in-variables" methods, to factor analysis and to principle component models. See e. g. [1] and the references given there. Here we want follow another line of thought. The disadvantage of the methods just mentioned is that they require the exogenous variables and the instruments to have a stochastic nature, usually even to be a drawing from a probability distribution which does not change over time. However this is really a crude way of "endogenizing" the exogenous variables and the instruments, which is in most cases not at all a realistic way of modelling these variables. Therefore we present here an alternative. The idea is simply to consider the *relations between the variables* as stochastic. This then leaves room for several of the variables to be chosen freely

(the exogenous variables and the instruments) and the remaining variables, the endogenous variables, are then determined by the stochastic relation(s); conditionally on the choice of the exogenous variables and the instruments the endogenous variables are stochastic. Which variables can serve as exogenous and instrument variables and which as endogenous variables follows from the estimation procedure, i. e. is determined a posteriori.

To enlighten these ideas we start with a simple case in section 2 in which this approach leads very naturally to the socalled *orthogonal regression.*

If one wants to apply such a scheme more generally, the question arises what class of probability measures one chooses on the set of stochastic relations. The problem here is that even for *linear* stochastic relations, the *set* of such relations is *not* a linear vector space. Therefore we study an analogon of the *mean* and the *variance* for general metric spaces, which we call the *centre(s)* and the *centre -variance.* And as an analogon of the Gaussian distribution on Euclidean space, we define the maximum entropy distribution, given the centre(s) and the centre-variance, on the metric space.

Next the sets of linear relations, the socalled Grassmannians, are considered. They can be handily represented by sets of symmetric idempotent matrices (also called orthogonal projection matrices) of a prescribed rank. A Grassmannian can be made into a metric space in several ways, one of which is just to take it to be equal to the metric space of symmetric idempotent matrices. Having done that we can calculate the maximum entropy distributions as described above and in the case of two variables present the maximum likelihood estimators that follow from this. We make some remarks about the maximum likelihood estimator in the general case and the paper finishes with some remarks, open questions and directions of further research.

2 A simple case

In order to present the idea of stochastic relations we start with a simple case. Consider the following regression model.

$$y_t = \beta x_t + u_t, \tag{1}$$

where β is a deterministic parameter (later on we will allow β to be stochastic) and u_t is stochastic, with a Gaussian distribution with mean α and variance σ^2 and with u_t and u_s stochastically independent if $t \neq s$.

This is the usual regression model, except for the fact that we have not yet made a statement about the nature of y_t, x_t. There are several possibilities:

(a) Both y_t and x_t are stochastic, with a parametrized probability distribution with known or unknown parameters. This leads to the errors-in-variables problem etc. [1].

(b) x_t is deterministic (or at least "predetermined", exogenous) and y_t is stochastic (endogenous). This leads to the standard least squares formulas of regression analysis.

(c) y_t is deterministic (or "predetermined", exogenous) and x_t is stochastic (endogenous). This leads again to the standard least squares formulas for regression analysis, but now with the role of y_t and x_t interchanged.

(d) Any of the preceding possibilities, but we do not know which one (or we do not want to use such information for one reason or another). In this case, which is the case that we want to treat in this paper, we will interpret an observation (x_t, y_t) as an observation on the stochastic relation that exists between x_t and y_t.

In the present case the relation between x_t and y_t that is imposed by the model (1) is a line in the (x, y) -plane. Its slope is determined by β, which is supposed to be deterministic in this case. However its constant term is stochastic, with Gaussian probability density, which makes that the set of parallel lines all with the same slope determined by β has a probability distribution. This probability distribution can be described in various ways, depending on how one parametrizes the set of parallel lines. It is crucial for the application of the Maximum Likelihood Principle to define a metric on each set of parallel lines. Because we only have parallel lines here, an obvious choice for the distance between any pair of lines is the minimal distance between any pair of points the first of which is from the first line and the second of which comes from the second line. Of course, this is the length of any piece of line which connects the two lines and is orthogonal to both of them. Any set of parallel lines can now be parametrized by choosing one of them (e.g. by the rule that it be the one that crosses the origin, or the centre of gravity of the observations) and desribing the others by their distance "to the right or to the left" of this first line. In this parametrizati on the probability distribution is again Gaussian, with as its mean the line $y = \beta x + \alpha$ and with variance $\frac{\sigma^2}{1+\beta^2}$.

Note that although we have implicitly excluded the vertical line in our formulation of the model (1) it would at this point be no problem to introduce it. In that case it would be natural to assume a Gaussian distribution over the set of verticals using the same concept of distance between two parallel lines as above. (This would not be possible if we measured the distance between two lines along the $y-$axis, i.e. by the distance between the two points where the two lines cross the $y-$axis.)

Before the Maximum Likelihood Estimator (MLE) can be derived a few remarks about its definition in this case should be made.

(1) In a metric space the following generalization of the Maximum Likelihood Estimator can be used. Let $d(\theta_1, \theta_2)$ denote the distance between two points θ_1, θ_2 in the metric space Θ. For each positive ϵ consider the $\epsilon-$balls $B(\theta_0, \epsilon) := \{\theta | d(\theta, \theta_0) < \epsilon\}$ and define the $\epsilon-$Maximum Likelihood Estimator as

$$\hat{\theta}_\epsilon = \arg\max_{\theta_0 \in \Theta} P[B(\theta_0, \epsilon)] \tag{2}$$

The (generalized) Maximum Likelihood Estimator $\hat{\theta}$ is now defined as the limit $\hat{\theta} = \lim_{\epsilon \downarrow 0} \hat{\theta}_\epsilon$. (In case $\hat{\theta}_\epsilon$ is a set of points one should take the limit of the sets $\hat{\theta}_\epsilon$, for $\epsilon \downarrow 0$.)

In case a probability density is defined on the metric space this implies the more usual notion that the estimator(s) can be found at the point(s) with maximal density. But even if the probability density has a support that is smaller than the whole of the metric space, this definition can still be applied. This definition also shows clearly the dependence of the M.L.E. on the metric that is used.

(2) Note that in our set-up we consider the observations as (sometimes partial) observations on the *relations*. Therefore the Maximum Likelihood Principle has to be applied to the probability density of (or at least-as treated in (1)- the probability of each small $\epsilon-$ball around) the (partially) observed relations, that are implied by the different models. This leads to the M.L.E. of the relation involved.

(3) In the present case we have not defined the distance between two non-parallel lines. In fact this is needed to apply (1) and (2). However the metric in the present case can be viewed as a limiting case of a generalization, involving scaling, of the metrics that will be defined in

section 4 on the set of *all lines* in $\mathbf{R}^2$ (and more generally on linear relations in $\mathbf{R}^n$). So the correct interpretation of the M.L.E. that will be derived in this section is as a limiting case of the M.L.E. that will be derived in the subsequent sections (if one allows for scaling).

In order to prepare for the following proposition, let us note some properties of the well-known orthogonal regression. Given T observations $(x_1, y_1), \ldots, (x_T, y_T)$, the orthogonal regression line is the one for which the sum of the squares of the distances of the points $(x_1, y_1), \ldots, (x_T, y_T)$ to that line is minimal over all possible lines (cf. e.g. [1]).

The orthogonal regression line can be found as follows. Let $b = (-\beta, 1)$, then the distance of a point $z_t = (x_t, y_t)$ to the line through the point $z_0 = (x_0, y_0)$ and orthogonal to the vector b is equal to $\frac{|b'(z_t - z_0)|}{\|b\|}$. The orthogonal regression line is obtained by minimization of the sum of squares of these distances, i. e. by minimization of $\sum_{t=1}^{T} \frac{(b'(z_t - z_0))^2}{b'b}$ with respect to b and z_0. Minimization with respect to z_0 can easily be seen to lead to choosing z_0 to be the centre of gravity: $z_0 = \bar{z} := \frac{1}{T} \sum_{t=1}^{T} z_t$

Let $\tilde{z}_t := z_t - z_0$. Then the minimization problem can be formulated as

$$\begin{aligned}
&\min_{b \neq 0} \sum_{t=1}^{T} \frac{(b'\tilde{z}_t)^2}{b'b} = \\
&\min_{b \neq 0} \sum_{t=1}^{T} tr\{(\frac{bb'}{b'b})\tilde{z}_t \tilde{z}_t'\} = \\
&\min_{b \neq 0} tr\{(\frac{bb'}{b'b}) \sum_{t=1}^{T} \tilde{z}_t \tilde{z}_t'\} = \\
&\min_{\Pi^2 = \Pi, rk\Pi = 1} T.tr(\Pi S), \qquad (3)
\end{aligned}$$

where the matrix Π stands for the rank one projection matrix $\frac{bb'}{b'b}$ and S denotes the sample covariance matrix of z. According to the theorem of Courant-Fischer the minimum is obtained by the projection matrix of rank one corresponding to the smallest eigenvalue of S.

We can now state the following result

Proposition 2.1 *The Maximum Likelihood Line Estimator, with respect to the described distance function between parallel lines, is equal to the line that results from orthogonal regression.*

Proof. For a given $b = (-\beta, 1)$ and z_0 the distance between the lines orthogonal to b and crossing the points z_t and z_0 respectively, is equal to $\frac{|b'(z_t - z_0)|}{\|b\|}$. I. e. it is equal to the distance of the point $z_t = (x_t, y_t)$ to the line through the point $z_0 = (x_0, y_0)$ and orthogonal to the vector b.

Because on the set of all lines orthogonal to b, parametrized with respect to these distances, the model (1) implies a Gaussian distribution, the M.L.E. is obtained by minimizing the sum of squares of these distances over the set of all lines orthogonal to b.

This implies that the M.L.E. is indeed the orthogonal regression line. If the variance of the Gaussian distribution is unknown, the sample mean of the squares of these distances is the M.L.E. of this variance.

Q.e.d.

3 Generalization of mean and variance to general metric spaces: Centre and centre-variance.

Use will be made of a generalization of mean and variance to spaces more general than Euclidean spaces. Generalizations of this type have been proposed under various headings by several authors. We mention e.g. [3]. Here we will use a definition for the case of metric spaces. Let Θ be a metric space with metric d as before.

Definition 3.1 *Let a probability measure P be defined on the metric space Θ and let E denote the corresponding expectation operator. For any point θ_0 consider the expectation $E[d(\theta, \theta_0)^2]$. Each point in Θ which minimizes this expression over all $\theta_0 \in \Theta$ is called a* centre *of the probability measure. The minimal value of $E[d(\theta, \theta_0)^2]$ is called the* centre-variance *or shortly the* variance *of the probability measure in the metric space.*

Remarks.

(1) A centre does not have to exist in general. However if the metric space is compact at least one centre exists. This follows from the continuity in this case of $E[d(\theta, \theta_0)^2]$ as a function of θ_0 , which in turn follows from the triangle inequality for the metric, together with the fact that in a compact metric space the metric is bounded.

A centre does not have to be unique. E. g. for the uniform distribution on the circle, each point on the circle is a centre.

(2) In general the variance does not have to be defined. If it exists it is unique by construction.

If the metric space is compact, the variance exists. This follows from the argument above.

(3) Clearly if the metric space is a Euclidean space with the standard metric, the centre is just the mean and the centre-variance is the usual variance.

(4) The concept of *unbiased estimator* can be defined in the same spirit. See [3].

(4) In the case of a von Mises distribution on the circle, the localization parameter is equal to the centre in our terminology. Compare Section 6

(5) The choice of the metric plays a crucial role. In many examples there are several "natural" metrics. E. g. in the case of a circle, the distance between two points can be measured by the angle between these two points ("the inner metric"); however one can also choose the chordal distance between those points ("outer metric").

For different choices of the metric different variance values and different centres may result.

4 A metric on the Grassmannians

A $k-$dimensional linear subspace S of the $n-$dimensional Euclidean space $\mathbf{R}^n$ can be represented in various ways, for example by a basis of k independent vectors from the subspace, or more generally as the image of some matrix Y with rank k. One possibility is to represent such a subspace by its orthogonal projection matrix (also called symmetric idempotent matrix in the econometric literature)

$$\Pi = \Pi(S) := Y(Y^TY)^\dagger Y^T, \tag{4}$$

where $\dagger$ denotes the operation of taking the Moore-Penrose generalized inverse of a matrix (see e. g. [4]); of course if the matrix is nonsingular, this coincides with the usual inverse. This representation is unique, i. e. independent of the specific choice of Y. This follows directly from the property of

such a matrix, that it maps each vector ξ in $\mathbf{R}^n$ to its orthogonal projection $\Pi\xi$ on the subspace involved. We will make use of this representation.

The set of all $k-$dimensional linear subspaces of the $n-$dimensional Euclidean space $\mathbf{R}^n$ is denoted by $G(k,n)$ and is called the Grassmannian manifold of $k-$planes in the $n-$dimensional Euclidean space. Using the representation of the elements of the Grassmannian by their orthogonal projection matrices, we will find it useful to consider $G(k,n)$ as the set of rank k orthogonal projection matrices of size $n \times n$. It is well-known that a Grassmannian is a differentiable manifold. It is in fact a Riemannian manifold, if one uses the standard Fubini-Study-Leichtweiss metric on the tangent bundle. The corresponding minimal arclength metric is the generalization of the angular distance on the circle ("inner metric"). A generalization of the chordal distance on the circle ("outer metric") is defined as follows:

Definition 4.1 *Let the distance function* $d : G(k,n) \times G(k,n) \longrightarrow [0,\infty)$ *be defined by* $d(\Pi_1,\Pi_2)^2 := tr[(\Pi_1 - \Pi_2)^2] = 2k - 2tr[\Pi_1\Pi_2]$, *i. e. the distance is equal to the socalled Frobenius norm of* $\Pi_1 - \Pi_2$.

This metric is the one that is induced by the standard Euclidean metric in $\mathbf{R}^{n\times n}$ by considering each orthogonal projection matrix Π in $G(k,n)$ as an n^2-vector in $\mathbf{R}^{n\times n}$. The representation of the elements of the Grassmannian by their orthogonal projection matrices produces therefore an isometric inbedding of the Grassmannian $G(k,n)$ in the n^2-dimensional Euclidean space $\mathbf{R}^{n\times n}$.

It forms clearly a closed and bounded subset of $\mathbf{R}^{n\times n}$ and therefore the well-known fact that a Grassmannian is compact follows.

5 Maximum entropy distributions on a Grassmannian

Combining the results of the previous two sections, it follows that any probability distribution on a Grassmannian has a finite variance and a well-defined centre or set of centres. In order to find an analogon of the Gaussian distribution in a general metric space, one can try to make use of the crucial property of a Gaussian distribution with mean μ and (scalar) variance σ^2 that it is the maximum entropy distribution with that mean and variance. In a general metric space the role of the mean is taken over by the centre and the role of the variance by the centre-variance. Therefore the problem

arises to find, given a point $\Pi_0 \in G(k,n)$ and a positive number σ^2, the maximum entropy distribution with centre Π_0 and centre-variance σ^2. We assert that the density $p = p(\Pi)$ of such a distribution (w.r.t. the volume element $dm(\Pi)$ that is derived from the Fubini-Study-Leichtweiss metric) takes the following form:

$$\begin{aligned} p(\Pi) &= \exp[-\frac{1}{2}\kappa d(\Pi,\Pi_0)^2 + \gamma] = \\ &= \exp[-\kappa\{k - tr(\Pi\Pi_0)\} + \gamma] \end{aligned} \tag{5}$$

where κ is a concentration parameter, related in a bijective way to the variance and γ is the normalization parameter. That Π_0 is the centre of the distribution in the case $\kappa \geq 0$, is formulated, among other related properties, in the next theorem.

Theorem 5.1 *Consider a probability density on $G(k,n)$ of the form (5). The expectation of Π with respect to this probability density is of the form*

$$E(\Pi) = c_0\Pi_0 + c_1(I - \Pi_0) \tag{6}$$

where c_0 and c_1 are nonnegative numbers.

If $c_0 > c_1$ then $\Pi_0 \in G(k,n)$ is the (unique) centre of the probability distribution. This occurs if $\kappa > 0$.

If $c_0 = c_1$ then the distribution is the uniform distribution on the Grassmannian, which corresponds to $\kappa = 0$.

If $c_0 \geq c_1$ the numbers c_0 and c_1 are related to the variance σ^2 of the distribution as follows:

$$\begin{aligned} c_0 &= (1 - \frac{\sigma^2}{2k}) \\ c_1 &= \frac{\sigma^2}{2(n-k)} \end{aligned} \tag{7}$$

Remark Note the difference between the *centre* of the distribution and the expectation $E(\Pi)$ which is well defined only due to our inbedding of the Grassmannian in $\mathbf{R}^{n\times n}$.

Proof Without loss of generality one can make an orthonormal change of basis such that $\Pi_0 \in G(k,n)$ takes the form

$\Pi_0 = \mathrm{diag}[1,\ldots,1,0,\ldots,0]$.

First it will be shown that $E(\Pi)$ is diagonal. Consider the set S of 2^n diagonal sign matrices $S = \mathrm{diag}[s_1,s_2,\ldots,s_n]$ with $s_i \in \{+1,-1\}$ for each

value of $i \in \{1,\ldots\}$. Given an arbitrary orthogonal projection matrix Π, and an arbitrary sign matrix $S \in \mathcal{S}$, the matrix $S\Pi S$ is again an orthogonal projection matrix. Furthermore due to the special form of Π_0 it is easy to see that $d(\Pi_0, S\Pi S) = d(\Pi_0, \Pi)$ for all $S \in \mathcal{S}$. Therefore for a fixed Π the density at all points $S\Pi S, S \in \mathcal{S}$ is equal. From this one obtains $E(\Pi) = E(S\Pi S) = SE(\Pi)S$ for each $S \in \mathcal{S}$ which in turn implies that $E(\Pi)$ is diagonal.

Let $\{e_1,\ldots,e_n\}$ denote the standard basis in $\mathbf{R}^n$. Let $\mathcal{T}$ denote the set of matrices which are obtained from the identity matrix by interchanging the i–th and j–th column, or equivalently, interchanging the i–th and j–th row. We will formally allow $i = j$ to hold, in which case no interchange of columns takes place. Let $\mathcal{T}_k \times \mathcal{T}_{n-k}$ denote the set of matrices which are obtained from the identity matrix by interchanging at most two of the first k columns and interchanging at most two of the last $n-k$ columns. Note that if $T \in \mathcal{T}_k \times \mathcal{T}_{n-k}$, then $T^2 = I$ and T is symmetric. Therefore, if Π is an orthogonal projection matrix, then so is $T\Pi T$. It can easily be seen that for an arbitrary $\Pi \in G(k,n)$ and any $T \in \mathcal{T}_k \times \mathcal{T}_{n-k}$ one has

$$d(\Pi_0, \Pi) = d(\Pi_0, T\Pi T). \tag{8}$$

From this it follows that for all $T \in \mathcal{T}_k \times \mathcal{T}_{n-k}$ one has

$$E(\Pi) = E(T\Pi T) = TE(\Pi)T \tag{9}$$

from which one can derive

$$\begin{aligned} E(\Pi) &= c_0 \mathrm{diag}[1,\ldots,1,0,\ldots,0] + c_1 \mathrm{diag}[0,\ldots,0,1,\ldots,1] = \\ &= c_0\Pi_0 + c_1(I - \Pi_0) \end{aligned} \tag{10}$$

This shows (6). Note that

$$\begin{aligned} \mathrm{tr}E(\Pi) &= E\mathrm{tr}\Pi = k = \\ &= c_0 \mathrm{tr}\Pi_0 + c_1 \mathrm{tr}(I - \Pi) = c_0 k + c_1(n-k). \end{aligned} \tag{11}$$

Suppose $c_0 > c_1$. In order to show that Π_0 is the unique centre of the distribution, we have to prove $E\|\Pi - \Pi_0\|_F^2 \leq E\|\Pi - \Pi_1\|_F^2$ for all $\Pi_1 \in G(k,n)$ and equality holds if and only if $\Pi_1 = \Pi_0$. This is equivalent to $\mathrm{tr}E(\Pi\Pi_0) \geq \mathrm{tr}E(\Pi\Pi_1)$ for all $\Pi_1 \in G(k,n)$ and equality holds if and only if $\Pi_1 = \Pi_0$.

From (6) it follows that

$$\begin{aligned} \mathrm{tr}[E(\Pi)\Pi_1] &= \mathrm{tr}[c_0\Pi_0\Pi_1 + c_1(I-\Pi_0)\Pi_1] = \\ &= \mathrm{tr}[(c_0-c_1)\Pi_0\Pi_1] + c_1\mathrm{tr}\Pi_1 = \\ &= (c_0-c_1)\mathrm{tr}[\Pi_0\Pi_1] + c_1 k. \end{aligned} \tag{12}$$

Now $\mathrm{tr}[\Pi_0\Pi_1]$ can be interpreted as the inner product $< \Pi_0, \Pi_1 >$ which corresponds to the Frobenius norm of matrices. Using this norm $\|\Pi\|^2 = k$ for all $\Pi \in G(k,n)$ and therefore $\frac{1}{k}\mathrm{tr}[\Pi_0\Pi_1] =< \frac{\Pi_0}{\sqrt{k}}, \frac{\Pi_1}{\sqrt{k}} >\leq 1$ and equality holds if and only if $\Pi_0 = \Pi_1$. This shows that indeed Π_0 is the unique centre if $c_0 > c_1$.

If $c_0 = c_1$ the same argument shows that $E\|\Pi - \Pi_0\|^2 = E\|\Pi - \Pi_1\|^2$ for all $\Pi_1 \in G(k,n)$ and therefore all elements of $G(k,n)$ are centres in this case.

So if $c_0 \geq c_1$ we know that Π_0 is a centre. Therefore the centre-variance is

$$\begin{aligned} \sigma^2 &= E(\|\Pi - \Pi_0\|^2) = \\ &= 2k - 2\mathrm{tr}[E(\Pi)\Pi_0] = \\ &= 2k - 2\mathrm{tr}[(c_0\Pi_0 + c_1(I-\Pi_0))\Pi_0] = \\ &= 2k - 2c_0 k = 2k(1-c_0). \end{aligned} \tag{13}$$

From this together with (11) it follows that

$$\begin{aligned} c_0 &= (1 - \frac{\sigma^2}{2k}) \\ c_1 &= \frac{\sigma^2}{2(n-k)} \end{aligned} \tag{14}$$

q.e.d.

Remark Without going into the proof (which is not difficult) let us state what the centres are in the case $\kappa < 0$, which is equivalent to $c_0 < c_1$. Three possibilities have to be distinguished:

(i) If $k < n-k$ then (the orthogonal projection operator of) each k–dimensional linear subspace of the $n-k$ dimensional image space of $I - \Pi_0$ is a centre.

(ii) If $k = n-k$ then $I - \Pi_0$ is the centre.

(iii) If $k > n-k$ then (the orthogonal projection operator of) each k–dimensional linear subspace which contains the $n-k$ dimensional image space of $I - \Pi$ is a centre.

Now we come to a sketch of the proof that the distributions given in (5) are indeed the maximum entropy distributions for a given centre Π_0 and a given centre-variance σ^2. Let $f(\Pi) := \log p(\Pi)$ denote the logarithm of a positive probability density $p(\Pi)$ on $G(k,n)$. The desired maximum entropy distribution is found by maximizing the entropy integral

$$\int_{G(k,n)} f(\Pi)\exp(f(\Pi))dm(\Pi) \tag{15}$$

under the restrictions

$$\begin{aligned} (1) \quad & \sigma^2 = E\|\Pi - \Pi_0\|^2 \Longleftrightarrow \\ & \mathrm{tr}[\int_{G(k,n)} \Pi\exp(f(\Pi))dm(\Pi)\Pi_0] = k - \frac{1}{2}\sigma^2 \\ (2) \quad & \int_{G(k,n)} \exp(f(\Pi))dm(\Pi) = 1. \end{aligned} \tag{16}$$

Note that we have *not* included the restriction that Π_0 is the centre! Therefore (1) should be interpreted as stating that the "Π_0−variance" $E\|\Pi-\Pi_0\|_F^2$ is equal to σ^2. Of course the centre-variance is by definition smaller than or equal to the Π_0−variance. Maximization of the entropy under (1) and (2) will turn out to lead to maximization of the centre-variance, and the maximal centre-variance is obtained if the centre-variance is equal to the Π_0−variance in which case Π_0 is by definition a centre.

So consider the Lagrangian

$$\begin{aligned} L(f) \;=\; & \int_{G(k,n)} f(\Pi)\exp(f(\Pi))dm(\Pi) + \\ & \lambda\{\mathrm{tr}[\int_{G(k,n)} \Pi\exp(f(\Pi))dm(\Pi)\Pi_0] + \\ & -(k - \frac{1}{2}\sigma^2)\} + \\ & +\mu\{\int_{G(k,n)} \exp(f(\Pi))dm(\Pi) - 1\} \end{aligned} \tag{17}$$

The function f is a stationary point of L if a variation δf of f produces a vanishing variation in the value of L :

$$\begin{aligned} 0 \;=\; & \frac{\delta L}{\delta f} = \\ =\; & \exp(f(\Pi)) + f(\Pi)\exp(f(\Pi)) + \\ & +\lambda\mathrm{tr}[\Pi\exp(f(\Pi))\Pi_0] + \\ & +\mu\exp(f(\Pi)) \end{aligned} \tag{18}$$

for all $\Pi \in G(k,n)$, which implies

$$f(\Pi) = -\lambda \mathrm{tr}[\Pi\Pi_0] - \mu - 1 \tag{19}$$

which shows that, with the correct choice for κ and γ the probability distribution takes the form (5) indeed.

Q.e.d.

6 The Von Mises distribution on the projective line

On the circle one of the standard probability distributions is the Von Mises distribution, given by the density

$$p(\theta) = \exp[\tilde{\kappa}\cos(\theta - \theta_0) + \tilde{\gamma}], \tag{20}$$

where θ_0 is the localization parameter, κ the concentration parameter and γ the normalization parameter; $\theta \in (0, 2\pi]$ the angle describing a point on a circle around the origin.

It is well-known that the (real) projective line, i.e. $G(1,2)$, is topologically equivalent to the circle. If we represent an element of $G(1,2)$, i. e. a line in $\mathbf{R}^2$ through the origin, by the angle $\omega \in (0,\pi]$ that the line will make with the $x-$axis, then a homeomorphism of the projective line to the circle is given by

$$\theta = 2\omega. \tag{21}$$

Using this homeomorphism, the maximum entropy distribution on $G(1,2)$ with given centre Π_0 and centre-variance σ^2 induces also a density on the circle. We will show that this density is in fact the Von Mises density.

A point in $G(1,2)$ is represented by a rank-one orthogonal projection matrix $\Pi = \frac{yy^T}{y^T y}, y \neq 0$. Let $\Pi_0 = \frac{y_0 y_0^T}{y_0^T y_0}, y_0 \neq 0$. Then the density on $G(1,2)$ is of the form

$$\begin{aligned} p(\Pi) &= \exp(\kappa \mathrm{tr}[(\frac{yy^T}{y^T y})(\frac{y_0 y_0^T}{y_0^T y_0})] + \gamma - k\kappa) = \\ &= \exp(\kappa[(\frac{y}{\|y\|})^T(\frac{y_0}{\|y_0\|})]^2 + \gamma - k\kappa) = \\ &= \exp(\kappa\cos^2(\omega - \omega_0) + \gamma - k\kappa), \end{aligned} \tag{22}$$

where $\omega - \omega_0$ is the angle between y and y_0. This is equal to

$$\begin{aligned} p(\Pi) &= \exp(\frac{\kappa}{2}\cos[2(\omega-\omega_0)] + \gamma - (k-\frac{1}{2})\kappa) = \\ &= \exp(-\tilde{\kappa}\cos(\theta-\theta_0) + \tilde{\gamma}) \end{aligned} \tag{23}$$

with $\tilde{\kappa} = \kappa/2$; $\theta = 2\omega$. The Riemannian volume element, which in this case is derived from the Fubini-Study metric on the projective line, produces a constant factor in the transformation to the circle; therefore only the normalization parameter $\tilde{\gamma}$ is affected by this and the form of the density remains the same. So indeed the Von Mises density is obtained this way.

7 The maximum likelihood estimator in the case of two variables

Consider the simple regression model (1), but now assume that β is stochastic and, for simplicity, $u_t = 0$ (Think for example of a model written in deviations from the mean). Thus for each $t \in \{0,\ldots,N\}$ the vector (y_t, x_t) lies on the line in $\mathbf{R}^2$ that is given by $y = \beta_t x$. Assume that this line is drawn at random from a probability distribution on $G(1,2)$ with density of the form (5). With each (nonzero) observation $(y_t, x_t) \neq (0,0)$ corresponds one possible line, namely the line in $G(1,2)$ that is represented by the orthogonal projection matrix $\Pi_t := Y_t(Y_t^T Y_t)^{-1} Y_t^T$ with $Y_t^T = (y_t, x_t)$.

Suppose one has N stochastically independent observations $\Pi_1,\ldots,\Pi_N$. The joint probability density $p(\Pi_1, \Pi_2, \ldots, \Pi_N)$ can easily be derived from (5) and $k = 1$ to be

$$p(\Pi_1,\ldots,\Pi_N) = \exp[\kappa \mathrm{tr}(\sum_{t=1}^{N} \Pi_t \Pi_0) + N(\gamma - \kappa)]. \tag{24}$$

So the maximum likelihood estimator $\hat{\Pi}$ of the *centre* is

$$\hat{\Pi} = \arg\max_{\Pi_0} \mathrm{tr}[(\sum_{t=1}^{N} \Pi_t)\Pi_0]. \tag{25}$$

This implies that the maximum likelihood estimator is the orthogonal projection matrix that corresponds with the eigenvector of the largest eigenvalue of $\sum \Pi_t$, if this eigenvalue has multiplicity one. In case of higher multiplicity, any one-dimensional linear subspace of the corresponding eigenspace is a maximum likelihood estimator (so in that case the MLE is not unique). For the variance the following result holds

Lemma 7.1 *The maximum likelihood estimator $\hat{\sigma}^2$ of the variance is*

$$\hat{\sigma}^2 = 2(1-\rho_1), \tag{26}$$

where ρ_1 is the largest eigenvalue of $\frac{1}{N}\sum_{t=1}^{N}\Pi_t$

Proof Consider the loglikelihood divided by N (recall that $k=1$):

$$\frac{1}{N}\log p(\Pi,\ldots,\Pi_N) = -\frac{1}{2}\kappa\{2-2\mathrm{tr}[(\frac{\sum_{t=1}^{N}\Pi_t}{N})\Pi_0]+\gamma(\kappa)\}, \tag{27}$$

where the fact that γ depends on κ is made explicit by writing γ as a function of κ. Substituting the maximum likelihood estimator $\hat{\Pi}_0$ for Π_0 and maximizing the likelihood with respect to κ one obtains the following first order condition

$$\begin{aligned} 0 &= \frac{\partial}{\partial\kappa}\log p(\Pi,\ldots,\Pi_N) = \\ &= -1+\mathrm{tr}[(\frac{\sum_{t=1}^{N}\Pi_t}{N})\Pi_0]+\frac{\partial\gamma}{\partial\kappa}. \end{aligned} \tag{28}$$

The derivative of γ with respect to κ can be calculated as follows:

Because γ is the normalization parameter it easily follows that

$$e^{-\gamma} = \int_{G(1,2)}\exp\{-\kappa(1-\mathrm{tr}[\Pi\Pi_0])\}dm(\Pi). \tag{29}$$

Differentiation with respect to κ gives

$$-\frac{\partial\gamma}{\partial\kappa} = -E(1-\mathrm{tr}[\Pi\Pi_0]) \tag{30}$$

wich is equal to $-\sigma^2/2$. So

$$\frac{\partial\gamma}{\partial\kappa} = \sigma^2/2 \tag{31}$$

and therefore the first order condition (28) leads to the following formula for the maximum likelihood estimator of the variance

$$\begin{aligned} \hat{\sigma}^2 &= 2-2\mathrm{tr}[(\frac{\sum_{t=1}^{N}\Pi_t}{N})\hat{\Pi}_0] = \\ &= 2(1-\rho_1) \end{aligned} \tag{32}$$

q.e.d.

Remark If one considers linear models with more than two variables and one or more simultaneous linear equations, the calculation of the M.L.E. becomes apparently more complicated. The reason is that in the general case the model describes for each value of t a $k-$dimensional linear subspace (with $k > 1$ in general) in which the $t-$th observation lies. Now given an observation there are lot of $k-$dimensional linear subspaces that contain that observation! So all one observes is an *event* in the sense of probability theory. (This is somewhat similar to throwing a die and observing not the exact number of spots up, but only that the number of spots up is even.) To calculate the likelihood one has to integrate the density over the event set, which leads to some complicated integrals. This subject needs further research.

8 Conclusions and remarks on possible further research

In this paper a set-up has been proposed to deal with the problem that, on various grounds, one does not always want to make the distinction between endogenous variables and exogenous variables a priori, i.e. before the estimation of the model. The way in which the problem is dealt with is to consider the *linear relations* themselves as stochastic. This makes it possible to consider a number of variables as free to choose; after such a choice the remaining variables are determined by the stochastic model and are therefore themselves stochastic. Use has been made of the maximum entropy distribution on a Grassmannian, given the centre, which is a generalization of the concept of a mean to a general metric space and given the centre-variance, which is a generalization of the concept of variance to a general metric space. By representing linear subspaces by their orthogonal projection matrices we were able to derive a number of results on the maximum entropy distributions. These were in turn used to study the maximum likelihood estimators. Only for the case of two variables an explicit expression for the MLE was presented. Research on the general case is still in progress. Let us make a number of final remarks about open problems and possibilities for further research.

(1) In this paper no attention has been given to scaling parameters, and this would certainly be an important next step; in fact that should give the analogon for this case of the usual variance-covariance matrix.

(2) Further research is needed to calculate the normalization parameters of the maximal entropy distributions on a Grassmannian.

(3) It is certainly possible to include a constant term in the model, in fact one can just apply the usual trick of introducing a dummy variable which has only one possible value, namely 1.

(4) Generalization to linear dynamical models is an interesting open problem.

References

[1] T. W. Anderson, *Estimating linear statistical relationships*, The Annals of Statistics, 1984, **12**, No.1, pp. 1-45.

[2] B. Hanzon, *Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems*, CWI Tracts 63,64, CWI, Amsterdam, 1989.

[3] H. Hendriks, *A Cramer-Rao type lower bound for estimators with values in a manifold*, Report 9015, Dept. Math., Univ. Nijmegen, Holland, March 1990.

[4] P. Lancaster, M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, 1985.

[5] R. E. Kalman, *Identifiability and Modeling in Econometrics*, in: P. S. Krishnaiah(ed.), *Developments in Statistics*,Acad. Press, New York, pp.97-136, 1983.

[6] J. C. Willems, *System Theoretical Models for the Analysis of Physical Systems*, Ricerche di Automatica, 1979, **10**, No. 2, pp. 71-106.

[7] J. C. Willems, *From time series to linear system. Part I: Finite dimensional linear time invariant systems.* Automatica, 1986, **22**,pp.561-580.

[8] J. C. Willems, *Paradigms and Puzzles in the Theory of Dynamical Systems*, IEEE Trans.Aut. Control, **36**, March 1991, pp. 259-294.

1990-1	B. Vogelvang	Testing For Co-Integration with Spot Prices of Some Related Agricultural Commodities
1990-2	J.C.J.M. van den Bergh P. Nijkamp	Ecologically Sustainable Economic Development Concepts and Model Implications
1990-3	J.C.J.M. van den Bergh P. Nijkamp	Ecologically Sustainable Economic Development in a Regional System: A Case Study in Agricultural Development Planning in the Netherlands
1990-4	C.Gorter P.Nijkamp P.Rietveld	Employers' Recruitment Behaviour and Re-Employment Probabilities of Unemployed
1990-5	K.Burger	Off-farm income and the farm-household the case of Kenyan smallholders
1990-6	H. Visser	Crowding out and the Government Budget
1990-7	P. Rietveld	Ordinal Data in Multicriteria Decision Making, a Stochastic Dominance Approach to Siting Nuclear Power Plants
1990-8	G. van der Laan P.H.M. Ruys D.J.J. Talman	Signaling devices for the supply of semi-public goods
1990-9	P.A.G. den Butter	Labour Productivity Slowdown and Technical Progress: An empirical analysis for The Netherlands
1990-10	R.W. van Zijp	Neo-Austrian Business Cycle Theory
1990-11	J.C. van Ours	Matching Unemployment and Vacancies: The Efficiency of the Dutch Labour Market
1990-12	B. Vogelvang	Hypotheses Testing Concerning Relationships between Spot Prices of Various Types of Coffee
1990-13	A.F. de Vos I.J. Steyn	Stochastic Nonlinearity: A Firm Basis for the Flexible Functional Form
1990-14	Y.H. van Emmerik D. de Jong W.W.A. Zunnmond D.N. Dukkers-van Emden	Opereren in overleg: geprotocolleerde samenwerking 1e-2e-lijn bij dagchirurgie
1990-15	T.J.J.R. Wolters	Mediation and Collective Bargaining: A Diagnostic Approach
1990-16	E.M.A. Scholten J. Koelewijn	Financieringsproblematiek van startende ondernemingen: een mogelijke verklaring op basis van empirisch onderzoek.
1990-17	E. Hfner H.P. Smit	Saturation and Model Specification of Passenger car Ownership
1990-18	P.A.G. den Butter	Sociale zekerheid, de wig en economische groei
1990-19	P.A.G. den Butter R.F. v.d. Wijngaert	Who is Correcting the Error ? A Co-integration Approach for Wages, Wage Space and Labour Conflicts in the Netherlands
1990-20	J.P. de Groot R. Ruben	Sistemas de Producción y Transferencia de Tecnología en la Economía Cafetalera de Centroamerica
1990-21	R. Ruben	Campesinado y Reforma Agraria en El Salvador
1990-22	J. van Ours G. Ridder	Vacancies and the Recruitment of New Employees
1990-23	A.F. de Vos J.J. de Vries	The Likelihood Function of a Generalized Gravity Model; Handling the Implicit Singularity of a Nonlinear Transformation
1990-24	D. van der Wal	Geloofwaardigheid en het Nederlandse wisselkoersbeleid
1990-25	R.J. Veldwijk M. Boogaard M.V. van Dijk E.R.K. Spoor	EDSOs, implosion and explosion: concepts to automate a part of application maintenance
1990-26	B. Hanzon	The area enclosed by the (oriented) Nyquist diagram and the Hilbert-Schmidt-Hankel norm of a linear system
1990-27	R.W. van Zijp	Why Lucas is not a Hayekian
1990-28	J. Rouwendal	On discrete choice under uncertainty: A generalization of the logit model and its application
1990-29	J. Rouwendal	On the equitable distribution of the housing stock
1990-30	J. Rouwendal	Stochastic market equilibria with efficient rationing (with an application to the Dutch housing market)
1990-31	J.A. Vijlbrief	The effects of unemployment insurance on the labour market
1990-32	J.G.W. Simons H.P. Wansink	Traffic Ban, a means to combat smog?
1990-33	J.C. van Ours T. Zoethout	De interne Arbeidsmarkt van de Gemeente Amsterdam
1990-34	H.J. Bierens	A Note On the Limiting Distribution of Sample Autocorrelations in the Presence of a Unit Root
1990-35	T. Kuhlman	The Economic Integration of Refugees in Developing Countries: A Research Model.
1990-36	T. Kuhlman	Towards a Definition of Refugees.
1990-37	T. Kuhlman	Organized versus Spontaneous Settlement of Refugees in Africa.
1990-38	R. Zuidema	The Neo-Austrian View on Interest
1990-39	G.v.d.Laan	General Equilibrium in a Closed International Trade Model