Title
The Dual of the Maximum Likelihood Method

Author
Paris, Quirino

Publication Date
2012-05-01

Department of Agricultural and Resource Economics
University of California, Davis

The Dual of the Maximum Likelihood Method

by

Quirino Paris

Working Paper No. 12-002

2012

Giannini Foundation of Agricultural Economics

The Dual of the Maximum Likelihood Method

Quirino Paris*

University of California, Davis

Summary

The Maximum Likelihood method estimates the parameter values of a statistical model that maximize the corresponding likelihood function, given the sample information. This is the primal approach that, in this paper, is presented as a mathematical programming specification whose solution requires the formulation of a Lagrange problem. A remarkable result of this setup is that the Lagrange multipliers associated with the linear statistical model (regarded as a set of constraints) is equal to the vector of residuals scaled by the variance of those residuals. The novel contribution of this paper consists in developing the dual model of the Maximum Likelihood method under normality assumptions. This model minimizes a function of the variance of the error terms subject to orthogonality conditions between the errors and the space of explanatory variables. An intuitive interpretation of the dual problem appeals to basic elements of information theory and establishes that the dual maximizes the net value of the sample information. This paper presents the dual ML model for a single regression and for a system of seemingly unrelated regressions.

Key Words: Maximum likelihood, primal, dual, value of sample information, SUR.

JEL Classification: C1, C3

* Quirino Paris is a professor at the University of California, Davis. May 2012.

1. Introduction

In general, to any problem stated in the form of a maximization (minimization) criterion, there corresponds a dual specification in the form of a minimization (maximization) goal. This structure must apply also to the Maximum Likelihood (ML) approach, one of the most powerful and widely used statistical methodologies. An intuitive description of the Maximum Likelihood principle is found in Kmenta (2011, pp. 175-180): Given a sample of observations about random variables, "the question is: To which population does the sample most likely belong?" The answer can be found by defining "the likelihood function as the joint probability distribution of the data, treated as a function of the unknown coefficients. The Maximum Likelihood Estimator (MLE) of the unknown coefficients consists of the values of the coefficients that maximize the likelihood function." (Stock and Watson, 2011). That is, the maximum likelihood estimator selects the parameter values that give the observed data sample the largest possible probability of having been drawn from a population defined by the estimated coefficients.

In this paper, we concentrate on data samples drawn from normal populations and deal with linear statistical models. The ML method, then, is to estimate the values of the mean and variance of that normal population that will maximize the likelihood function given the sample information.

The novel contribution of the paper consists in developing the dual specification of the Maximum Likelihood method (under normality) and in giving it an intuitive interpretation. In section 2 we present the traditional Maximum Likelihood method in the form of a primal nonlinear programming model. This specification is a natural step toward the discovery of the dual structure of the Maximum Likelihood method. It differs from the traditional way to deal with the ML approach because the estimates of parameters and errors are computed simultaneously rather then sequentially. A remarkable result of this setup is that the Lagrange multipliers associated with the linear statistical model (regarded as a set of constraints) is equal to the vector of residuals scaled by the variance of those residuals. Section 3 derives and interprets the dual specification of the ML method. It turns out that a crucial dual constraint is represented by a non-traditional definition of the variance σ^2 as the inner product of the vector of sample information y and the vector of errors u divided by the number of observations. The intuitive interpretation of the dual problem appeals to the basic elements of information theory and establishes that the dual maximizes the net value of the sample information (NVSI), as defined and illustrated in section 3. Section 4 presents a discussion of the dual specification of the seemingly unrelated regressions (SUR) model. Aside from the more complex structure of the SUR specification and of the corresponding likelihood function, the process of deriving the dual problem is analogous to the single equation methodology. Also the intuitive interpretation of the SUR dual appeals to basic elements of information theory and establishes the maximization of the net value of sample information. Although the numerical solution of the dual ML method is feasible (given a stable nonlinear programming software), the main goal of this paper is to present and interpret the dual specification of the ML method because it completes the understanding of the role played by residuals. It turns out, in fact, that residuals in the primal model assume the role of shadow prices in the dual model, as discussed in section 3. Section 5 suggests an implementable numerical specification of the dual SUR model.

2. The Primal of the Maximum Likelihood Method (Normal Distribution)

Consider a linear statistical model

$$y = X\beta + u \tag{1}$$

where y is an $(n \times 1)$ vector of sample data (observations), X is a $(n \times k)$ matrix of predetermined values, β is an $(k \times 1)$ vector of unknown parameters, and u is an $(n \times 1)$ vector of random errors that is assumed to be independently and normally distributed as $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $\sigma^2 > 0$. The vector of observations y is known also as the sample information. Then, ML estimates of the parameters and errors can be obtained by maximizing the logarithm of the corresponding likelihood function with respect to (u, β, σ^2) subject to the constraints represented by model (1). We called this specification the primal ML method. Specifically,

Primal: $$\max \log \ (u, \beta, \sigma^2 \mid y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{u'u}{2\sigma^2} \tag{2}$$

subject to $$y = X\beta + u. \tag{3}$$

Traditionally, the constraints (3) are substituted for the error vector u in the likelihood function (2) and an unconstrained maximization calculation will follow. This algebraic manipulation, however, obscures the path toward a dual specification of the problem under study. Therefore, we maintain the structure of the ML model as in relations (2) and (3) and proceed to state the corresponding Lagrangean function by selecting an $(n \times 1)$ vector ε of Lagrange multipliers of constraints (3) to obtain

$$L(u, \beta, \sigma^2, \varepsilon) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{u'u}{2\sigma^2} - \varepsilon'(y - X\beta - u). \tag{4}$$

Following Kmenta (2011, pp. 181-182), the maximization of the Lagrangean function (4) requires taking partial derivatives with respect to $(u, \beta, \sigma^2, \varepsilon)$, setting them equal to zero, in which case we signify that a solution of the resulting equations is an estimate of the corresponding parameters and errors:

$$\frac{\partial L}{\partial u} = -\frac{u}{\sigma^2} + \varepsilon \tag{5}$$

$$\frac{\partial L}{\partial \beta} = X'\varepsilon \tag{6}$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{u'u}{2\sigma^4} \tag{7}$$

$$\frac{\partial L}{\partial \varepsilon} = -y + X\beta + u \;. \tag{8}$$

First order necessary conditions (FONC) are given by equating (5)-(8) to zero and signifying that the resulting solution values are ML estimates. We obtain

$$\hat{\varepsilon} = \frac{\hat{u}}{\hat{\sigma}^2} \tag{9}$$

$$X'\hat{\varepsilon} = 0 \tag{10}$$

$$\hat{\sigma}^2 = \hat{u}'\hat{u} / n \tag{11}$$

$$y = X\hat{\beta} + \hat{u} \tag{12}$$

where $\hat{\sigma}^2 > 0$. A first remarkable observation is that the Lagrange multipliers $\hat{\varepsilon}$ are defined in terms of the estimates of the primal variables u and σ^2. Indeed, the Lagrange multipliers $\hat{\varepsilon}$ are equal to $\hat{u}$ up to a scalar $\hat{\sigma}^2$. This recognition will simplify the statement of the corresponding dual problem because we can dispense with the symbol ε.

In fact, equation (10) can be restated equivalently as the following orthogonality condition between the residuals of model (1) and the space of predetermined variables

$$X'\hat{\varepsilon} = X'\frac{\hat{u}}{\hat{\sigma}^2} = 0 . \tag{13}$$

The meaning of a Lagrange multiplier (or dual variable) is that of a marginal sacrifice (shadow price) which is a measure of tightness of the corresponding constraint. Within the context of this paper, the optimal values of the Lagrange multipliers $\hat{\varepsilon}$ are identically equal to the residuals $\hat{u}$ scaled by $\hat{\sigma}^2$. Hence, the symbol $\hat{u}$ takes on a double role: as a vector of residuals in the primal ML problem [(2)-(3)] and as a vector of shadow prices in the dual ML problem (defined in the next section).

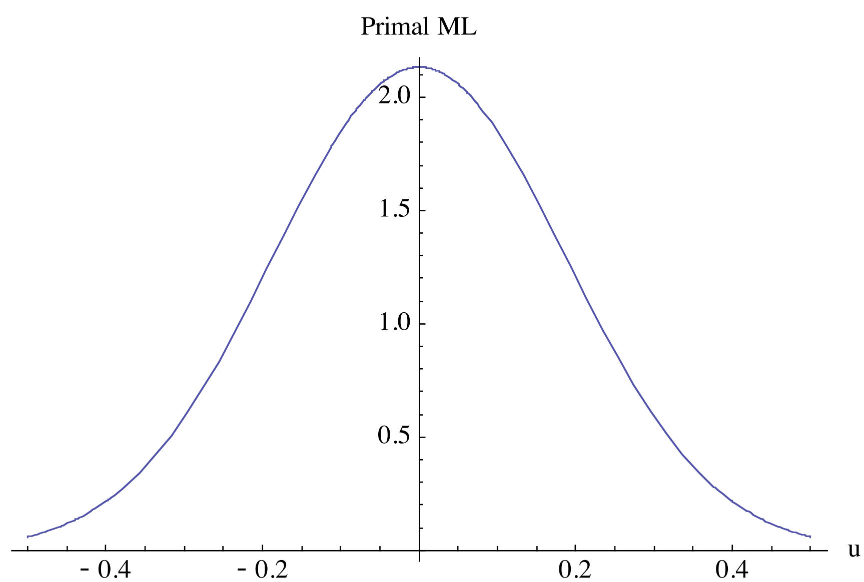

Figure 1. Representation of the antilog of the primal objective function (2) for one normal variable with zero mean and $\sigma^2 = 0.036$.

Figure 1 illustrates the familiar shape of the likelihood function – as represented in the primal objective function (2) – for one normal random variable.

Furthermore, note that the multiplication of equation (12) by the vector $\hat{\varepsilon}$ (replaced by its equivalent representation given in equation (9)) and the use of the orthogonality condition (13) produces

$$\begin{aligned} \hat{\varepsilon}' y &= \hat{\varepsilon}' X\hat{\beta} + \hat{\varepsilon}'\hat{u} \\ \frac{\hat{u}' y}{\hat{\sigma}^2} &= \frac{\hat{u}'\hat{u}}{\hat{\sigma}^2} \\ \hat{u}' y &= \hat{u}'\hat{u}. \end{aligned} \tag{14}$$

This means that a ML estimate of σ^2 can be obtained equivalently as

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n} = \frac{\hat{u}' y}{n}, \tag{15}$$

a computationally useful result. Relation (15), in fact, is of paramount importance for stating an operational specification of the dual ML model because the presence of the vector of sample observations y in the estimate of σ^2 is the only place where this sample information exercises its structural effects in the dual model by limiting the range of the residuals $\hat{u}$, as clarified in the next section.

3. The Dual of the Maximum Likelihood Method

The statement of the dual specification of the ML approach follows the general rules of duality theory in mathematical programming. The dual objective function to be minimized is the Lagrangean function stated in (4) with the appropriate simplifications allowed by its algebraic structure and the information derived from FONCs (9) and (10). The dual constraints are all the FONCs that are different from the primal constraints. This leads to the following specification:

$$\textbf{Dual:}\quad \min L^*(u,\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{u' y}{\sigma^2} + \frac{u' u}{2\sigma^2} \tag{16}$$

$$\text{subject to} \qquad \frac{X'u}{\sigma^2} = 0 \tag{17}$$

$$\sigma^2 = \frac{u'y}{n}\,. \tag{18}$$

The solution of the dual ML problem [(16)-(18)] by any reliable nonlinear programming software (e.g., GAMS) produces estimates of (u, β, σ^2) that are identical to those obtained from the primal ML problem. The dual ML estimate of the parameter vector β is obtained as a vector of Lagrange multipliers associated to the orthogonality constraints (17). The last two terms of the dual objective function can be rearranged to express a quantity that has intuitive appeal, that is

$$-\frac{u'y}{\sigma^2} + \frac{u'u}{2\sigma^2} = -\frac{(u'y - u'u/2)}{\sigma^2}\,. \tag{19}$$

Using simple terminology of information theory, the expression in the numerator of the second ratio of equation (19) can be given the following interpretation. The linear statistical model (1) can be regarded as the decomposition of a message into a signal and noise, that is

$$\text{message} = \text{signal} + \text{noise} \tag{20}$$

where $\text{message} \equiv y$, $\text{signal} \equiv X\beta$, and $\text{noise} \equiv u$. Let us recall that the double role of u includes also that of shadow price of the constraints defined by model (1). Hence, the inner product $u'y$ can be regarded as the gross value of the sample information while the expression $u'u/2$ can be regarded as a cost function of noise. In summary, therefore, the relation $(u'y - u'u/2)$ can be interpreted as the net value of sample information (*NVSI*) that is maximized in the dual objective function.

It is well known that, under the assumptions of model (1), least-squares estimates of the parameters are also maximum likelihood estimates. Hence, $u'u/2$ is the primal objective function (to be minimized) of the least-squares method while $(u'y - u'u/2)$ corresponds to the dual objective function (to be maximized) of the least-squares approach. An optimal solution of either the primal or the dual ML problems results in

$$\min_u LS = \frac{u'u}{2} = u'y - \frac{u'u}{2} = \max_u NVSI$$
$$\hat{LS} = \frac{\hat{u}'\hat{u}}{2} = \hat{u}'y - \frac{\hat{u}'\hat{u}}{2} = \hat{N}VSI$$

where LS stands for Least Squares. The role of shadow price taken on by $\hat{u}$ is justified by the derivative of the least-squares function with respect to the vector of sample information y, that is

$$\frac{\partial \hat{LS}}{\partial y} = \hat{u}$$

indicating that the change in the Least-Squares objective function due to an infinitesimal change in the vector of constraints (sample information in model (1)) is the meaning of the Lagrange multipliers that, in the LS case, are equal to the vector of residuals.

Note also that, in view of constraint (18), relation (19) can be reformulated as

$$-\frac{u'y}{\sigma^2} + \frac{u'u}{2\sigma^2} = -n + \frac{u'u}{2\sigma^2} \qquad (21)$$

a result that justifies the use of the alternative definition of the variance σ^2 in constraint (18), that follows from equation (14). That definition, in fact, provides the only effective presence of the sample information y in the dual model.

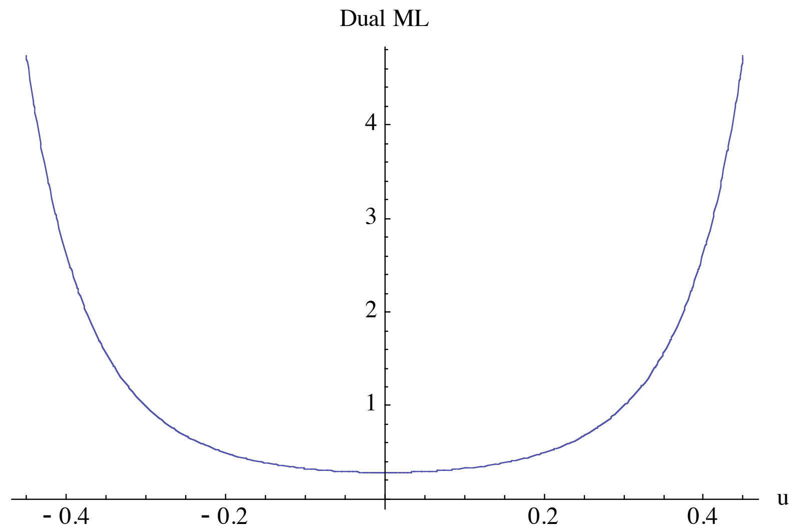

Figure 2. Representation of the antilog of the dual objective function (16) for one normal variable with zero mean and $\sigma^2 = 0.036$.

4. The Dual of the Seemingly Unrelated Regressions Model

A seemingly unrelated regressions model consists of M linear equations defined by a sample of T observations for each equation;

$$\mathbf{y}_m = X_m \boldsymbol{\beta}_m + \mathbf{u}_m \tag{22}$$

where $m = 1,...,M$, $\mathbf{y}_m$ is a $(T \times 1)$ vector of sample observations, X_m is an $(T \times K_m)$ matrix of predetermined values, $\boldsymbol{\beta}_m$ is a $(K_m \times 1)$ vector of regression coefficients, and $\mathbf{u}_m$ is a $(T \times 1)$ vector of disturbances. We assume that the disturbances $\mathbf{u}_m$ are normally distributed with mean zero, $E(u_{mt}) = 0$, $t = 1,...,T$ and covariance matrix $E(\mathbf{u}_m \mathbf{u}'_m) = \sigma_{mm} \mathbf{I}_T$ where $\mathbf{I}_T$ is an identity matrix of order $(T \times T)$. Furthermore, we assume that the disturbances in different equations are contemporaneously correlated, that is $E(\mathbf{u}_m \mathbf{u}'_p) = \sigma_{mp} \mathbf{I}_T$, $(m, p = 1,...,M)$.

In order to simplify the notation and to facilitate the derivatives of the likelihood function we write model (22) in a more compact fashion, as in Schmidt (1976)

$$\begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \vdots & \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \vdots & X_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 & 0 & \vdots & 0 \\ 0 & \boldsymbol{\beta}_2 & \vdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \vdots & \boldsymbol{\beta}_M \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \vdots & \mathbf{u}_M \end{bmatrix} \quad (23)$$

and, finally:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \mathbf{U} \quad (24)$$

where $\mathbf{Y}$ is a $(T \times M)$ matrix, $\mathbf{X}$ is a $(T \times \sum_m^M K_m)$ matrix, $\boldsymbol{B}$ is a $(\sum_m^M K_m \times M)$ matrix, and $\mathbf{U}$ is a $(T \times M)$ matrix. Given the assumptions, the error covariance is an $(M \times M)$ matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2M} \\ \cdots & \cdots & & \cdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sigma_{MM} \end{bmatrix}. \quad (25)$$

The logarithm of the likelihood function of the above SUR specification constitutes the primal objective function to be maximized subject to the linear system of constraints stated in equation (24). Analytically, the primal problem becomes

Primal: $$\max \log \ell(\mathbf{U}, \boldsymbol{B}, \Sigma) = -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log|\Sigma| - \frac{1}{2} Trace\,\mathbf{U}\Sigma^{-1}\mathbf{U}' \quad (26)$$

$$\text{subject to} \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{B} + \mathbf{U}\,. \quad (27)$$

The interpretation of the primal problem corresponds to the traditional maximum likelihood approach where the goal is to select those values of the parameters $(\boldsymbol{B}, \Sigma)$ that maximize the probability of the sample $\mathbf{Y}$. In the above nontraditional specification of the ML approach, the interpretation of the primal problem should be integrated with the statement that the goal is to select those values of matrices ($\mathbf{U}, \boldsymbol{B}, \Sigma$) that maximize the probability of the sample $\mathbf{Y}$.

The path toward the dual of problem [(26),(27)] begins with the statement of the Lagrangean function and the calculation of its first derivatives. We select a ($T \times M$) matrix of Lagrange multipliers $\boldsymbol{\Lambda}$ associated with constraints (27) and write

$$\begin{aligned} L(\mathbf{U}, \boldsymbol{B}, \Sigma, \boldsymbol{\Lambda}) &= -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log|\Sigma| - \frac{1}{2} Trace\, \mathbf{U}\Sigma^{-1}\mathbf{U}' - Trace\, \boldsymbol{\Lambda}'(\mathbf{Y} - \mathbf{X}\boldsymbol{B} - \mathbf{U}) \\ &= -\frac{MT}{2}\log(2\pi) + \frac{T}{2}\log|\Sigma^{-1}| - \frac{1}{2} Trace\, \Sigma^{-1}\mathbf{U}'\mathbf{U} - Trace\, \boldsymbol{\Lambda}'(\mathbf{Y} - \mathbf{X}\boldsymbol{B} - \mathbf{U}) \end{aligned} \tag{28}$$

with first partial derivatives

$$\frac{\partial L}{\partial \mathbf{U}} = -\mathbf{U}\Sigma^{-1} + \boldsymbol{\Lambda} \tag{29}$$

$$\frac{\partial L}{\partial \boldsymbol{B}} = \mathbf{X}'\boldsymbol{\Lambda} \tag{30}$$

$$\frac{\partial L}{\partial \Sigma^{-1}} = \frac{T}{2}\Sigma - \frac{1}{2}\mathbf{U}'\mathbf{U} \tag{31}$$

$$\frac{\partial L}{\partial \boldsymbol{\Lambda}} = -\mathbf{Y} + \mathbf{X}\boldsymbol{B} + \mathbf{U}\,. \tag{32}$$

MLE of the unknown parameters are obtained by equating relations (29)-(32) to zero and solving for $(\tilde{\mathbf{U}}, \tilde{\boldsymbol{B}}, \tilde{\Sigma}, \tilde{\boldsymbol{\Lambda}})$

$$\tilde{\boldsymbol{\Lambda}} = \tilde{\mathbf{U}}\tilde{\Sigma}^{-1} \tag{33}$$

$$\mathbf{0} = \mathbf{X}'\tilde{\boldsymbol{\Lambda}} \tag{34}$$

$$\tilde{\Sigma} = \frac{\tilde{\mathbf{U}}'\tilde{\mathbf{U}}}{T} \tag{35}$$

$$\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{B}} + \tilde{\mathbf{U}}\,. \tag{36}$$

The dual specification takes advantage of the structure of the Lagrangean function (28) and the FONCs (33), (34) and (36). Note that, using (33) and (34), and multiplying (36) by the matrix of Lagrange multipliers $\tilde{\Lambda}$, we obtain

$$\begin{aligned} \tilde{\boldsymbol{\Lambda}}'\mathbf{Y} &= \tilde{\boldsymbol{\Lambda}}'\mathbf{X}\tilde{\boldsymbol{B}} + \tilde{\boldsymbol{\Lambda}}'\tilde{\mathbf{U}} \\ \tilde{\Sigma}^{-1}\tilde{\mathbf{U}}'\mathbf{Y} &= \tilde{\Sigma}^{-1}\tilde{\mathbf{U}}'\tilde{\mathbf{U}} \end{aligned} \tag{37}$$

and, therefore,

$$\begin{aligned} Trace\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}'\mathbf{Y} &= Trace\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}'\tilde{\mathbf{U}} \\ Trace\mathbf{Y}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' &= Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' \\ Trace\tilde{\boldsymbol{\Lambda}}'\mathbf{Y} &= Trace\mathbf{Y}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' \\ Trace\tilde{\Lambda}'\tilde{\mathbf{U}} &= Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' \end{aligned} \tag{38}$$

Hence, the Lagrangean function (28) can be evaluated and simplified as

$$\begin{aligned} L(\tilde{\mathbf{U}},\tilde{\Sigma}) &= -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log\left|\tilde{\Sigma}\right| - \frac{1}{2}Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' - Trace\tilde{\boldsymbol{\Lambda}}'\mathbf{Y} + Trace\tilde{\boldsymbol{\Lambda}}'\mathbf{X}\tilde{\boldsymbol{B}} + Trace\tilde{\boldsymbol{\Lambda}}'\tilde{\mathbf{U}} \\ &= -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log\left|\tilde{\Sigma}\right| - \frac{1}{2}Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' - Trace\mathbf{Y}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' + Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' \\ &= -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log\left|\tilde{\Sigma}\right| - Trace\mathbf{Y}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' + \frac{1}{2}Trace\tilde{\mathbf{U}}\tilde{\Sigma}^{-1}\tilde{\mathbf{U}}' \end{aligned}$$

This leads to the statement of the dual specification of the seemingly unrelated regressions model as

Dual: $$\min L(\mathbf{U},\Sigma) = -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log\left|\Sigma\right| - Trace\mathbf{Y}\Sigma^{-1}\mathbf{U}' + \frac{1}{2}Trace\mathbf{U}\Sigma^{-1}\mathbf{U}' \tag{39}$$

subject to $$\mathbf{0} = \mathbf{X}'\mathbf{U}\Sigma^{-1} \tag{40}$$

$$\Sigma = \frac{\mathbf{U}'\mathbf{U}}{T}. \tag{41}$$

The dual constraints (40) and (41) are a generalization of constraints (17) and (18). Note that, from equation (37), the error covariance matrix could have been stated, in analogy to constraint (18), as

$$\Sigma = \frac{\mathbf{U}'\mathbf{Y}}{T}. \tag{42}$$

This is, in fact, a more stable specification of the covariance matrix for the actual computation of the ML estimates using the dual approach. The estimate of the parameter matrix $\boldsymbol{B}$ is obtained as the matrix of Lagrange multipliers of the orthogonality constraints (40). Since Lagrange multipliers have the meaning of shadow prices and $\tilde{\mathbf{\Lambda}} = \tilde{\mathbf{U}}\tilde{\Sigma}^{-1}$, the last two terms of the dual objective function (39) can be interpreted as the net value of the sample information (*NVSI*) (analogous to the single equation example), with $Trace\mathbf{Y}\Sigma^{-1}\mathbf{U}'$ being the gross value of sample information and $Trace\mathbf{U}\Sigma^{-1}\mathbf{U}'/2$ the cost of noise:

$$NVSI = Trace\mathbf{Y}\Sigma^{-1}\mathbf{U}' - Trace\mathbf{U}\Sigma^{-1}\mathbf{U}'/2. \tag{43}$$

This quantity is maximized in the dual problem due to the negative sign appearing in front of it.

5. Numerical Implementation of the Dual SUR Model

The main goal of this paper was aimed at the analysis and intuitive understanding of the other (dual) side of the ML method. It did not consist in the proposal that ML estimates be obtained by solving the dual model numerically. This goal, however, can easily be achieved with a stable nonlinear software such as GAMS. At present time, traditional econometric packages cannot be used for this task because they do not deal explicitly with the error structure of econometric models.

The principal step toward a numerical implementation of the dual ML SUR model is the acknowledgement of two facts: first, it is necessary to define explicitly the determinant of the Σ matrix. Second, it is necessary to compute the inverse of Σ that must be guaranteed to be symmetric and positive definite. The determinant of a square matrix A can be computed by appealing to its Cholesky factorization $A = LDL^T$ where L is a unit lower triangular matrix, L^T is its transpose, and D is a diagonal matrix with positive elements on the main diagonal. It follows that the determinant of matrix A is the product of the diagonal elements of the D matrix. The inverse of matrix A is computable by stating explicitly the relation $AA^{-1} = I$. Then, a proper coding for GAMS of the dual ML SUR problem (39)-(41) can be stated as

$$\min \mathbf{L}(\bullet) = -\frac{MT}{2} - \frac{T}{2}\log[det(\Sigma)] - \sum_{t-1}^{T}\sum_{m=1}^{M}\sum_{p=1}^{M} y_{tm}\sigma^{mp}u_{tp} + \sum_{t-1}^{T}\sum_{m=1}^{M}\sum_{p=1}^{M} u_{tm}\sigma^{mp}u_{tp} / 2 \qquad (44)$$

$$\text{subject to} \qquad \sum_{p=1}^{M}\sum_{t=1}^{T} x_{tk}u_{tp}\sigma^{pm} = 0, \qquad m = 1,...,M; k = 1,....\sum_{k=1}^{M} K_m \qquad (45)$$

$$\Sigma = \left[\sum_{t=1}^{T} u_{tm}y_{tp} / T\right], \qquad m, p = 1,...,M \qquad (46)$$

$$\Sigma = LDL^T \qquad (47)$$

$$det(\Sigma) = \prod_{m=1}^{M} d_{mm} \qquad (48)$$

$$\Sigma\Sigma^{-1} = I \qquad (49)$$

where σ^{mp} is the *m*th and *p*th element of the inverse matrix Σ^{-1}. The ML estimate of the matrix of parameters $\boldsymbol{B}$ is obtained as a matrix of Lagrange multipliers associated with the system of constraints (45). The Cholesky factorization of constraint (47) defines the symmetry and the positive definiteness of the covariance matrix Σ. The diagonal matrix D is required to have positive elements on its main diagonal. A very important aspect of solving highly nonlinear models by numerical methods is a proper scaling of the data series. As a rule of thumb, a desirable process is to scale all the data series around a unit value. The computation of the covariance matrix as in constraint (46) is required by the fact that it is the only place where the sample information appears in the dual model. This is because the objective function term involving the vector of sample information y corresponds to

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{p=1}^{M} y_{tm}\sigma^{mp}u_{tp} = TM \; . \tag{50}$$

Equation (50) is nothing else that the multivariate extension of equation (21) for the single variable case. This dual ML formulation was solved for a SUR linear model (with intercepts) whose sample data included four dependent variables and five explanatory variables.

6. Conclusion

A statistical model may be intuitively regarded as the decomposition of a sample of messages into signal and noise. When the noise is distributed according to a normal density, the ML method maximizes the probability that the sample belong to a given normal population defined by the ML estimates of the model's parameters. All this is well known. This paper has analyzed and interpreted the dual of the ML method under univariate and multivariate normal assumptions.

It turns out that the dual function is a convex function of noise. Hence, a convenient interpretation is that the dual of the ML method minimizes a cost function of noise. This cost function is defined by the sample variance (in the univariate case) or the sample covariance matrix (in the multivariate case) of noise. Equivalently, the dual ML method maximizes the net value of the sample information. The choice of an economic terminology for interpreting the dual ML method is justified by the double role played by the symbols representing the residual terms of the estimated statistical model. These symbols may be interpreted as residual quantities in the primal specification and shadow prices in the dual model because they correspond to the Lagrange multipliers of the sample observations.

Under the assumption of this paper, the least-squares method leads to ML estimates of the model's parameters. Therefore, also in this case, the dual of the least-squares approach corresponds to the maximization of the net value of the sample information.

The numerical implementation of the dual ML method brings to the fore, by necessity, a neglected definition of the sample variance (in the univariate case) and the sample covariance matrix (in the multivariate case). In both cases, it is necessary to state the definition of the variance and covariances as the product of the sample information and the residuals (noise), as revealed by equations (15) and (42), because it is the only place where the sample information appears in the dual model. The dual approach to ML estimates may or may not be of any practical importance over the traditional estimation procedures. This depends on the researcher's personal preferences for numerical algorithms. It completes, however, our knowledge of what exists on the other side of the ML fence, a territory that has revealed interesting vistas on old statistical landscapes.

References

Benoit, Commandant. (1924), "Notes sur une Méthode de Résolution des Équations Normales Provenant de l'Application de la Méthode des Moindres Carrés a un Systéme d'Équations Linéaires en Nombre Inférieur a Celui des Inconnues.-

Application de la Méthode a la Résolution d'un Systéme Defini d'Équations Linéaires, (Procédé du Commandant Cholesky)." *Bull. Géodésique* 2(April, May, June 1924):67-77.

Brooke, A., Kendrick, D. and Meeraus, A. (1988), GAMS – A User's Guide, Redwood City, CA: The Scientific Press.

Kmenta, J. (2011), Elements of Econometrics – Second Edition, Ann Arbor: The University of Michigan Press.

Schmidt, P. (1976), Econometrics, New York: Marcel Dekker.

Stock, J. H. and Watson, M. W. (2011), Introduction to Econometrics – Third Edition, Boston: Addison-Wesley.