AN APPLICATION OF THE FRANK-WOLFE ALGORITHM AT MAXIMUM LIKELIHOOD ESTIMATION PROBLEMS

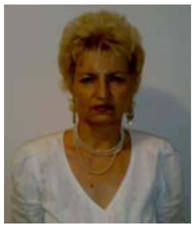
Ciprian Costin POPESCU

PhD Candidate, University Assistant
Department of Mathematics, University of Economics, Bucharest, Romania
Calea Dorobantilor 15-17 Street, Sector 1, Bucharest, Romania

E-mail: cippx@yahoo.com

Lia POPESCU

Department of Mathematics, Economic College
Rosiori de Vede, Romania



Abstract: *This paper tackles the problem of maximum likelihood estimation [2] under various types of constraints (equalities and inequalities restrictions) on parameters. The initial model, which is in fact a maximization problem (here are a few methods available in literature for estimating the parameters: ERM (expectation-restricted-maximization) algorithms, GP (gradient projection) algorithms and so on) is change into a new problem, a minimization problem. This second form is suited to a variant of Frank-Wolfe method for solving linearly restricted nonlinear programming problems [5]. In this way, some difficulties from the previous approaches are removed.*

Key words: *Constrained maximum likelihood; Nonlinear programming; Frank-Wolfe algorithm*

Developments and algorithms[1]

There are many situations in statistical computation which implies maximum likelihood estimation. The aim of this work is to generalize a model developed by Jamshidian [2] (by introducing a supplementary inequality constraint at the left) and to solve them using a regularization of FW-algorithm [5]. Thus, consider the optimization problem:

$$\max l(\boldsymbol{\theta})$$

subject to
$$\begin{cases} \mathbf{a}_i^T \boldsymbol{\theta} = b_i, i \in I_1 \\ b_i^- \le \mathbf{a}_i^T \boldsymbol{\theta} \le b_i^+, i \in I_2 \end{cases} \qquad (1)$$

where $\boldsymbol{\theta} = (\theta_1,...,\theta_p) \in R^p, \boldsymbol{\theta} \geq \mathbf{0}$, $\mathbf{a}_i = (a_{i_1},...,a_{i_p}) \in R^p$, $I_1 = \{i_1,...,i_m\}$ and $I_2 = \{i_{m+1},...,i_{m+n}\}$, $\mathrm{card}I_1 = m$, $\mathrm{card}I_2 = n$.

From (1) we get:

$$\max l(\boldsymbol{\theta})$$

subject to
$$\begin{cases} \mathbf{a}_i^T\boldsymbol{\theta} = b_i, i \in I_1 \\ \mathbf{a}_i^T\boldsymbol{\theta} \leq b_i^+, i \in I_2 \\ -\mathbf{a}_i^T\boldsymbol{\theta} \leq -b_i^-, i \in I_2 \end{cases} \tag{2}$$

Now we denote

$$\begin{cases} \mathbf{a}_{i_1} = \mathbf{u}_{i_1} \\ \vdots \\ \mathbf{a}_{i_m} = \mathbf{u}_{i_m} \end{cases}, \begin{cases} b_{i_1} = v_{i_1} \\ \vdots \\ b_{i_m} = v_{i_m} \end{cases}, \begin{cases} \mathbf{a}_{i_{m+1}} = \mathbf{u}_{i_{m+1}} \\ \vdots \\ \mathbf{a}_{i_{m+n}} = \mathbf{u}_{i_{m+n}} \end{cases}, \begin{cases} b^+_{i_{m+1}} = v_{i_{m+1}} \\ \vdots \\ b^+_{i_{m+n}} = v_{i_{m+n}} \end{cases} \tag{3}$$

and

$$\begin{cases} -\mathbf{a}_{i_{m+1}} = \mathbf{u}_{i_{m+n+1}} \\ \vdots \\ -\mathbf{a}_{i_{m+n}} = \mathbf{u}_{i_{m+2n}} \end{cases}, \begin{cases} b^-_{i_{m+1}} = v_{i_{m+n+1}} \\ \vdots \\ b^-_{i_{m+n}} = v_{i_{m+2n}} \end{cases} \tag{4}$$

Let $I_3 = \{i_{m+1},...,i_{m+n},...,i_{m+2n}\}$.
The problem (2) is equivalent with:

$$\max l(\boldsymbol{\theta})$$

s.t.
$$\begin{cases} \mathbf{u}_i^T\boldsymbol{\theta} = v_i, i \in I_1 \\ \mathbf{u}_i^T\boldsymbol{\theta} \leq v_i, i \in I_3 \end{cases} \tag{5}$$

or

$$\max l(\boldsymbol{\theta})$$

s.t.
$$\begin{cases} \mathbf{u}_{i_k}^T\boldsymbol{\theta} = v_{i_k}, k = \overline{1,m} \\ \mathbf{u}_{i_k}^T\boldsymbol{\theta} \leq v_{i_k}, k = \overline{m+1, m+2n} \end{cases} \tag{6}$$

Under the assumptions that $\theta_1 = \mu_1,...,\theta_p = \mu_p$ and $f : R^{p+2n} \to R, f(\boldsymbol{\mu}) = (l(\boldsymbol{\theta}) + 0\cdot\mu_{p+1} + ... + 0\cdot\mu_{p+2n})$ where $\boldsymbol{\mu} \in R^{p+2n}, \mu = (\boldsymbol{\theta}, \mu_{p+1},...,\mu_{p+2n}) = (\theta_1,...,\theta_p,\mu_{p+1},...,\mu_{p+2n}) =$
$= (\mu_1,...,\mu_p,\mu_{p+1},...,\mu_{p+2n})$
we may formulate (6) as:

$$\begin{aligned} &\min f(\boldsymbol{\mu}) \\ &\text{s.t.} \begin{cases} \mathbf{u}_{i_k}^T \boldsymbol{\theta} = v_{i_k}, k = \overline{1,m} \\ \mathbf{u}_{i_k}^T \boldsymbol{\theta} + \mu_{p+k-m} = v_{i_k}, k = \overline{m+1, m+2n} \end{cases} \end{aligned} \tag{7}$$

In the matricial form we have:

$$\begin{aligned} &\min f(\boldsymbol{\mu}) \\ &\text{s.t.} \begin{pmatrix} \mathbf{u}_{i_1}^T & 0 & \dots & & 0 \\ \vdots & \vdots & & & \vdots \\ \mathbf{u}_{i_m}^T & 0 & 0 & \dots & 0 \\ \mathbf{u}_{i_{m+1}}^T & 1 & 0 & \dots & 0 \\ \mathbf{u}_{i_{m+2}}^T & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ \mathbf{u}_{m+2n-1}^T & 0 & \dots & 1 & 0 \\ \mathbf{u}_{i_{m+2n}}^T & 0 & \dots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \\ \\ \vdots \\ \vdots \\ \\ \\ \mu_{p+2n} \end{pmatrix} = \begin{pmatrix} v_{i_1} \\ \\ \\ \vdots \\ \vdots \\ \\ \\ v_{i_{m+2n}} \end{pmatrix} \end{aligned} \tag{8}$$

or

$$\begin{aligned} &\min f(\boldsymbol{\mu}) \\ &\begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{U}_2 & \mathbf{I}_{2n} \end{pmatrix} \boldsymbol{\mu} = \mathbf{v} \end{aligned} \tag{9}$$

where

$$\mathbf{U}_1 = \begin{pmatrix} \mathbf{u}_{i_1}^T \\ \vdots \\ \mathbf{u}_{i_m}^T \end{pmatrix}, \ \mathbf{U}_2 = \begin{pmatrix} \mathbf{u}_{i_{m+1}}^T \\ \vdots \\ \mathbf{u}_{i_{m+2n}}^T \end{pmatrix}.$$

If

$$\mathbf{A} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{U}_2 & \mathbf{I}_{2n} \end{pmatrix}$$

then (9) is equivalent with

$$\begin{aligned} &\min f(\boldsymbol{\mu}) \\ &\text{s.t.} \\ &\mathbf{A}\boldsymbol{\mu} = \mathbf{v} \end{aligned} \tag{10}$$

where

$\boldsymbol{\mu} \in R^{p+2n}, \mathbf{A} \in M_{m+2n,p+2n}(R), \mathbf{v} \in R^{m+2n}$

or

$$\begin{aligned}&\min f(\boldsymbol{\mu})\\&\text{s.t.}\\&\boldsymbol{\mu} \in X\end{aligned} \qquad (11)$$

where

$$X = \{\mathbf{x} \in R^{p+2n} / \mathbf{Ax} = \mathbf{v}, \mathbf{x} \geq \mathbf{0}\}$$

Discussion and conclusion

The problem in the form (11) is suitable for applying a variant of Frank-Wolfe method (the regularized algorithm-RFW) (see Migdalas [5]):

For $\boldsymbol{\mu}^k \in X$, the objective function f is approximated by $\nabla f(\boldsymbol{\mu}^k)^T \boldsymbol{\mu}$ and (11) becomes:

$$\begin{aligned}&\min \nabla f(\boldsymbol{\mu}^k)^T \boldsymbol{\mu}\\&\text{s.t.}\\&\boldsymbol{\mu} \in X\end{aligned} \qquad (12)$$

The regularization of the problem means that an additional term appears in the objective function such that the distance between the iteration point $\boldsymbol{\mu}^k$ and the solution $\tilde{\boldsymbol{\mu}}^k$ is restricted. It is proved [2] that the point $\boldsymbol{\mu}^k$ is a solution for (11) if and only if it verifies the regularized subproblem:

$$\begin{aligned}&\min \nabla f(\boldsymbol{\mu}^k)^T \boldsymbol{\mu} + t_k \phi(\boldsymbol{\mu}, \boldsymbol{\mu}^k)\\&\text{s.t.}\\&\boldsymbol{\mu} \in X\end{aligned} \qquad (13)$$

Moreover, the regularized Frank-Wolfe algorithm, given below, is convergent [4,5].

-Step 1: consider $\boldsymbol{\mu}^0 \in X, t_0 = t > 0, k = 0$.

-Step 2: consider $\boldsymbol{\mu}^k$ the solution for (11) and let $\mathbf{d}^k = \tilde{\boldsymbol{\mu}}^k - \boldsymbol{\mu}^k$. If $\mathbf{d}^k = \mathbf{0}$, stop.

-Step 3: for $\tilde{\alpha}^k = \max\{\alpha / \boldsymbol{\mu}^k + \alpha^k \mathbf{d}^k\} \in X$ seek after $\alpha^k \in \operatorname{argmin}\{f(\boldsymbol{\mu}^k + \alpha^k \mathbf{d}^k), \alpha \in [0, \tilde{\alpha}^k]\}$. Let $\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \alpha^k \mathbf{d}^k, t_{k+1} = t_k, k = k+1$. Go to step 2.

Rererences

1. Frank, M. Wolfe, P. **An algorithm for quadratic programming,** Naval Research Logistic Quarterly, Vol. 3, 1956, pp. 95-110
2. Jamshidian, M. **On algorithms for restricted maximum likelihood estimation,** Computational Statistics & Data Analysis, Vol. 45, 2004, pp. 135-157
3. Jamshidian, M. Jennrich, R. I. **Acceleration of the EM algorithm by quasi-Newton methods,** Journal of Royal Statistical Society, Vol 59, 1997, pp. 569-587

4. Martos, B. **Nonlinear Programming: Theory and Methods,** North-Holland, Amsterdam, 1975
5. Migdalas, A. **A regularization of the Frank-Wolfe method and unification of certain nonlinear programming methods,** Mathematical Programming, Vol. 65, 1994, pp. 331-345
6. Wolfe, P. **Convergence theory in nonlinear programming,** J. Abadie, ed., Integer and Nonlinear Programming, North-Holland, Amsterdam, 1970, pp. 1-36

[1] Codifications of references:

[1]	Frank, M. Wolfe, P. **An algorithm for quadratic programming,** Naval Research Logistic Quarterly, Vol. 3, 1956, pp. 95-110
[2]	Jamshidian, M. **On algorithms for restricted maximum likelihood estimation,** Computational Statistics & Data Analysis, Vol. 45, 2004, pp. 135-157
[3]	Jamshidian, M. Jennrich, R. I. **Acceleration of the EM algorithm by quasi-Newton methods,** Journal of Royal Statistical Society, Vol 59, 1997, pp. 569-587
[4]	Martos, B. **Nonlinear Programming: Theory and Methods,** North-Holland, Amsterdam, 1975
[5]	Migdalas, A. **A regularization of the Frank-Wolfe method and unification of certain nonlinear programming methods,** Mathematical Programming, Vol. 65, 1994, pp. 331-345
[6]	Wolfe, P. **Convergence theory in nonlinear programming,** J. Abadie, ed., Integer and Nonlinear Programming, North-Holland, Amsterdam, 1970, pp. 1-36