COPYRIGHT NOTICE

FedUni ResearchOnline
<http://researchonline.ballarat.edu.au>

This is the submitted for peer-review version of the following article:

Mammadov, M., Zhao, L., & Zhang, J., (2012). The effect of regularization on drug-reaction relationships. *Optimization*. 61(4). 405-422

Which has been published in final form at:

<http://dx.doi.org/10.1080/02331934.2011.641547>

© Taylor and Francis Group, LLC.

This is the author's version of the work. It is posted here with permission of the publisher for your personal use. No further distribution is permitted.

THE EFFECT OF REGULARIZATION ON DRUG-REACTION RELATIONSHIPS

M. Mammadov,^{a,b*} L. Zhao^a and J. Zhang^c

^a*University of Ballarat, Victoria 3353, Australia;* ^b*National Information and Communications Technology Australia (NICTA);* ^c*Shanghai University, Shanghai 200444, China*

The least squares method is a standard approach used in data fitting that have important applications in many areas in Science and Engineering including many finance problems. In the case when the problem under consideration involves large-scale sparse matrices regularization methods are used to obtain more stable solutions by relaxing the data fitting. In this paper, a new regularization algorithm is introduced based on the Karush-Kuhn-Tucker conditions and the Fisher-Burmeister function. The Newton method is used for solving corresponding systems of equations. The advantages of the proposed method has been demonstrated in the establishment of drug-reaction relationships based on the Australian Adverse Drug Reaction Advisory Committee (ADRAC) database.

Keywords: optimization, adverse drug reaction; data classification; regularization method; ill-posed problem

AMS Subject Classification: 90C30; 90C90

1. Introduction

The linear least squares fitting (LLSF) technique is the simplest and most commonly applied form of linear regression and provides a solution to the problem of finding the best fitting straight line through a set of points. This method has also been accepted as one of the most efficient data classification methods [27]. Its applications to Finance and different regularized versions are quite vast; see for example, pricing options ([2, 24, 26]) and eco-finance and gene-environment networks ([13, 20]).

We use the following notations. $A = (a_{ij})$, $i = 1, \dots, m, j = 1, \dots, n$, is a $(m \times n)$ matrix, where m is the number of samples and n is the number of features. We are interested in the case when the number m is much greater than n ; that is, $m \gg n$. The class information is given by a $(m \times c)$ matrix $B = (b_{ij})$, $i = 1, \dots, m, j = 1, \dots, c$, where c is the number of classes. In the applications below we assume that $b_{ij} = 1$ if sample i belongs to class j , and $b_{ij} = 0$ if not.

Consider the following problem

$$\text{minimize } \|AX - B\|^2, \text{ subject to } X \in R^{n \times c}, \quad (1)$$

where $\|\cdot\|$ denotes the Frobenius norm: $\|A\| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$. This problem is about finding a $(n \times c)$ matrix $X = (x_{ij})$, $i = 1, \dots, n, j = 1, \dots, c$, such the difference between matrices AX and B is minimal in terms of the Frobenius norm.

We present these matrices by column vectors $a_i = (a_{1i}, \dots, a_{mi})^T$, $i = 1, \dots, n$, $b_j = (b_{1j}, \dots, b_{mj})^T$ and $x_j = (x_{1j}, \dots, x_{nj})^T$, $j = 1, \dots, c$, in the form $A = (a_1, \dots, a_n)$, $B = (b_1, \dots, b_c)$ and $X = (x_1, \dots, x_c)$. Then problem (1) can be divided into c different independent problems. For each $j = 1, \dots, c$, we have the following problem

$$\text{minimize } \|Ax_j - b_j\|^2, \text{ subject to } x_j \in R^n. \quad (2)$$

It is clear that if x_j , $j = 1, \dots, c$, is a solution to (2) then the matrix $X = (x_1, \dots, x_c)$ is a solution to (1).

The theory of ill-posed problems has been well developed in the literature. Hadamard [6] defined a problem to be *ill-posed* if the solution is not unique or if an arbitrarily small perturbation of the data (noise from matrix B) can cause an arbitrarily large perturbation of the solution.

This situation encounters frequently when considering noisy data sets. In such cases highly perturbed solutions X to (1) may not present “good” *feature-class* (*drug-reaction* in this paper) relationships.

In order to obtain some better estimation to the unknown “true” solution $\hat{X}$, different regularization techniques have been developed. The main idea of regularization methods is to incorporate additional information about the unknown “true” solution $\hat{X}$ together with the data fitting; that is, minimization of $\|AX - B\|$. This strategy is referred to as *regularization* [7]. For example, the classical Tikhonov regularization [9, 11, 21, 25] applies the following optimization problem:

$$\text{minimize } \|AX - B\|^2 + \lambda \|X\|^2, \text{ subject to } X \in R^{n \times c}. \quad (3)$$

The philosophy of regularization is to find a balance between the data fitting $\|AX - B\|$ and regularization $\|X\|$. The regularization parameter λ in (3) is used for this aim. By minimizing $\|AX - B\|$, we can achieve a better data fitting, and by minimizing the norm of X we can find more “stable” solutions.

When dealing with this problem, a major difficulty is the selection of a “correct” choice of regularization parameter λ . In the literature many methods have been developed for this aim, however, the “correct” choice of regularization parameter λ in the Tikhonov regularization, is still an open question, especially for large scale problems.

An alternative to (3) might be the following two problems called Morozov’s discrepancy principle [19] and Ivanov’s regularization [12]:

$$\text{minimize } \|X\|^2, \text{ subject to } \|AX - B\| \leq \delta; \quad (4)$$

$$\text{minimize } \|AX - B\|^2, \text{ subject to } \|X\| \leq \epsilon. \quad (5)$$

It can be proved [4, 8] that problems (3), (4) and (5) are equivalent, via the Lagrange multipliers, provided that λ , ϵ and δ have been accordingly selected. However, (4) and (5) are more suitable in the applications as they allow for more flexibility in the choice of parameters ϵ and δ .

In this paper we consider data classification on, so called Short Featured Multi-Labeled (SFML) data sets from the Adverse Drug Reaction (ADR) problems. The terminology SFML was first used in [18] when studying text classification problems, however such data sets have been previously studied in ADR problems (see [17] and references therein). In this paper we show that these data sets lead to large-scale ill-posed problems and, therefore, regularization techniques are required for the development of LLSF type classification algorithms. The advantages of such classifiers are briefly discussed in Section 3 (see [17] for more details).

Through analyzing two factors $\|AX - B\|$ and $\|X\|$ for ADR problems we conclude that the relaxation of data fitting $\|AX - B\|$ in favor of minimizing $\|X\|$ is more preferable in terms of classification accuracy on new samples. Thus, the formulation (4) will be used in the development of a new classification algorithm. We introduce a new method for solving problem (4) by applying the Karush-Kuhn-Tucker conditions and the Fisher-Burmeister function. We use the Newton method for solving corresponding systems of equations.

For the purpose of comparison, we adopt two other regularization algorithms: the Conjugate Gradients Least Squares (CGLS) and the hybrid method with W-GCV regularization (HyBR). CGLS was first proposed by Hestenes and Stiefel [10], and further extended by A. Bjorck [1]. It is a useful method for large scale data sets. In connection with discrete ill-posed problems, it is an interesting fact that when the Conjugate Gradient algorithm is applied to the unregularized normal equations $A^T A \mathbf{x} = A^T \mathbf{b}$ then the low-frequency components of the solution tend to converge faster than the high-frequency components. Thus CGLS has some inherent regularization effects where the number of iterations plays the role of the regularization parameter. HyBR has been proposed recently by Chung et. al. [3] for solving large-scale ill-posed inverse problems. Because this method can choose a regularization parameter at each iteration appropriately, it is often adopted to overcome the semi-convergence behavior of the Lanczos method.

The rest of this paper is organized as follows. In the next section we present a new numerical method for solving regularization with discrepancy. Adverse Drug Reaction problems are considered in Section 3; in particular, it is demonstrated that they lead to ill-posed problems. In Section 4 we present the results of numerical experiments to illustrate the performance of the discussed regularization methods. Conclusions are given in Section 5.

2. A new numerical method for solving regularization with discrepancy

In this paper, we introduce a new regularization method for solving the following optimization problem motivated from (4):

$$\text{minimize } \frac{1}{2} \|Lx\|^2, \quad \text{subject to } \frac{1}{2} \|Ax - b\|^2 \leq \frac{1}{2} \delta^2. \quad (6)$$

In the theory of regularization methods, the $p \times n$ matrix L is usually the identity matrix or an approximation of the derivative operator. Similar to (2), this problem refers to considering an optimization problem for a particular class.

We will denote by Φ the feasible set in problem (6):

$$\Phi = \{x \in R^n : \|Ax - b\| \leq \delta\}.$$

Clearly, Φ is convex; that is, given any $x_1, x_2 \in \Phi$ and $x_\mu = \mu x_1 + (1 - \mu)x_2$,

$\mu \in (0, 1)$, we have $x_\mu \in \Phi$:

$$\|Ax_\mu - b\| = \|\mu(Ax_1 - b) + (1 - \mu)(Ax_2 - b)\| \leq \mu\delta + (1 - \mu)\delta = \delta.$$

Let

$$l^* = \inf\{\|Lx\| : x \in \Phi\} \quad (7)$$

and S stand for the set of solutions to problem (6):

$$S = \{x \in \Phi : \|Lx\| = l^*\}. \quad (8)$$

In [14] a more general form of $\phi(x)$ for the objective function in problem (6) is considered. However, in the applications to large scale ill-posed problems it is more practical to use simpler forms, like $\phi(x) = \frac{1}{2}\|Lx\|^2$, or even $\phi(x) = \frac{1}{2}\|x\|^2$.

A new regularization technique based on truncated iterations of the Lagrangian method is introduced in [14] and [15]. The advantages of the formulation (6) are demonstrated in the applications to image denoising problems.

The following theorems below describe the structure of the solution set S and in particular establish the existence and uniqueness of the solution to problem (6).

Theorem 2.1 : *Let feasible set Φ be not empty. Then problem (6) has a solution; that is, $S \neq \emptyset$. Moreover, S is a closed and convex set.*

Proof: From continuity of functions $\|Lx\|$ and $\|Ax - b\|$ it follows that S is a closed set. Let $x_1, x_2 \in S$ and $x_\mu = \mu x_1 + (1 - \mu)x_2$, $\mu \in (0, 1)$. Since Φ is convex, we have $x_\mu \in \Phi$. On the other hand

$$\|Lx_\mu\| \leq \mu\|Lx_1\| + (1 - \mu)\|Lx_2\| = \mu l^* + (1 - \mu)l^* = l^*.$$

In view to (7) this means that $\|Lx_\mu\| = l^*$. Therefore $x_\mu \in S$; that is, S is a closed and convex set.

Now we prove that S is not empty.

If $\|b\| \leq \delta$ then $x = 0 \in S$ is a solution to problem (6). Consider the case

$$\|b\| > \delta. \quad (9)$$

Consider a sequence $x_k \in \Phi$ such that

$$\lim_{k \rightarrow \infty} \|Lx_k\| = l^*. \quad (10)$$

From $x_k \in \Phi$ it follows that $\|Ax_k\|$ is bounded; that is, there is a number $M < \infty$ such that

$$\|Ax_k\| \leq M, \quad \forall k. \quad (11)$$

1. If sequence x_k is bounded then there is a convergent subsequence that for the sake of simplicity will be denoted again by x_k . Let $x_k \rightarrow \tilde{x}$. Then we have $\|A\tilde{x} - b\| \leq \delta$ and $\|L\tilde{x}\| = l^*$. This means that $\tilde{x} \in S$ or $S \neq \emptyset$.

Consider the case when x_k is unbounded:

$$\lim_{k \rightarrow \infty} \|x_k\| = \infty. \quad (12)$$

In this case the sequence $x_k/\|x_k\|$ is bounded and therefore has a convergent subsequence. Again, for the simplicity we assume that

$$\frac{x_k}{\|x_k\|} \rightarrow e_1 \text{ as } k \rightarrow \infty.$$

Clearly $\|e_1\| = 1$. From (11), (12) we obtain

$$\|A \frac{x_k}{\|x_k\|}\| \leq \frac{M}{\|x_k\|} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

Therefore, $\|Ae_1\| = 0$; that is,

$$e_1 \in \text{Null}(A) \doteq \{x \in R^n : Ax = 0\}. \quad (13)$$

Now if $n = 1$ then we have $Ax = 0$ for all $x \in R^n$. Then, from (9) it follows that feasible set Φ is empty. This contradicts the assumptions of the theorem.

2. Let $n \geq 1$. Consider a sequence y_k defined by

$$y_k = x_k - \xi_k^1 e_1, \text{ where } \xi_k^1 = \langle x_k, e_1 \rangle.$$

Here $\langle \cdot, \cdot \rangle$ stands for the scalar product. From (10) and (12) we observe that

$$\|Le_1\| = \lim_{k \rightarrow \infty} \|L \frac{x_k}{\|x_k\|}\| = \lim_{k \rightarrow \infty} \frac{1}{\|x_k\|} \|Lx_k\| = 0. \quad (14)$$

Thus $e_1 \in \text{Null}(L)$. Then, we have:

- (1) $\langle y_k, e_1 \rangle = \langle x_k, e_1 \rangle - \xi_k^1 \langle e_1, e_1 \rangle = 0$;
- (2) $Ay_k = Ax_k - \xi_k^1 Ae_1 = Ax_k$; this in particular means that
 - $\|Ay_k - b\| \leq \delta$,
 - $\|Ay_k\| \leq M$;
- (3) $Ly_k = Lx_k - \xi_k^1 Le_1 = Lx_k$; this in particular means that
 - $\|Ly_k\| \rightarrow l^*$ as $k \rightarrow \infty$.

If sequence y_k is bounded then there is a convergent subsequence that for the simplicity will be denoted by y_k . Let $y_k \rightarrow \tilde{y}$. Then we have $\|A\tilde{x} - b\| \leq \delta$ and $\|L\tilde{x}\| = l^*$. This means that $\tilde{y} \in S$ and therefore $S \neq \emptyset$.

Consider the case when y_k is unbounded:

$$\lim_{k \rightarrow \infty} \|y_k\| = \infty. \quad (15)$$

Then, similar to part **1** above, we can consider a convergent subsequence

$$\frac{y_k}{\|y_k\|} \rightarrow e_2 \text{ as } k \rightarrow \infty,$$

where it is not difficult to see that $\|e_2\| = 1$ and

$$\langle e_1, e_2 \rangle = \lim_{k \rightarrow \infty} \frac{\langle e_1, y_k \rangle}{\|y_k\|} = 0.$$

Similar to (13) we can show that $e_2 \in \text{Null}(A)$; that is, $Ae_2 = 0$.

Now if $n = 2$ then each $x \in R^n$ can be represented as $x = \mu_1 e_1 + \mu_2 e_2$. In this case we have $Ax = \mu_1 Ae_1 + \mu_2 Ae_2 = 0$ for all $x \in R^n$. Then, from (9) it follows that feasible set Φ is empty. This contradicts the assumptions of the theorem.

3. If $n > 2$ we can continue the above procedure by considering a sequence y_k defined by

$$y_k = x_k - \xi_k^1 e_1 - \xi_k^2 e_2, \quad \text{where } \xi_k^1 = \langle x_k, e_1 \rangle, \quad \xi_k^2 = \langle x_k, e_2 \rangle.$$

If sequence y_k is bounded then we obtain a solution to problem (6) as a limit point of this sequence. Otherwise, repeating the steps in **2** we can have a contradiction for the case if $n = 3$.

This procedure with steps **1-3** can be continued at most n steps, which means that set S is not empty.

Theorem is proved. $\square$

Theorem 2.2: *Let feasible set Φ be not empty and the following condition hold:*

$$\text{Null}(L) \cap \text{Null}(A) = \{0\}. \quad (16)$$

Then the set S of solutions to problem (6) is bounded.

Proof: On the contrary assume that there is a sequence of solutions $x_k \in S$ such that

$$\|x_k\| \rightarrow \infty, \quad \|Ax_k - b\| \leq \delta \quad \text{and} \quad \|Lx_k\| = l^*.$$

For the sake of simplicity, let

$$\lim_{k \rightarrow \infty} \frac{x_k}{\|x_k\|} = e.$$

We have $\|e\| = 1$ and similar to (13), (14) we obtain that

$$e \in \text{Null}(L) \cap \text{Null}(A)$$

which contradicts (16).

Theorem is proved. $\square$

Theorem 2.3: *Assume that $L^T Lx \neq 0$ for all $x \in \Phi$. Let $\bar{x}$ be a solution to problem (6) and $\bar{\lambda}$ be a corresponding Lagrange multiplier. Then $\bar{\lambda} > 0$ and $\|A\bar{x} - b\| = \delta$.*

Proof: If $\|A\bar{x} - b\| < \delta$ then $\bar{x}$ is an interior point of feasible set Φ . Then the gradient of the function $\|Lx\|^2$ at this point should be zero:

$$\nabla\left(\frac{1}{2}\|L\bar{x}\|^2\right) = L^T L\bar{x} = 0.$$

This contradicts the assumption of the theorem. Thus, $\|A\bar{x} - b\| = \delta$.

Now we show that $\bar{\lambda} > 0$. From the first-order necessary conditions for (6) we have

$$\begin{cases} L^T L\bar{x} + \bar{\lambda} A^T (A\bar{x} - b) = 0, \\ \bar{\lambda} \geq 0, \quad \delta^2 \geq \|A\bar{x} - b\|^2, \\ \bar{\lambda}(\delta^2 - \|A\bar{x} - b\|^2) = 0. \end{cases} \quad (17)$$

If $\bar{\lambda} = 0$, then from the first two equations of (17), $L^T L\bar{x} = 0$, and $\|A\bar{x} - b\|^2 \leq \delta^2$, which is a contradiction.

Theorem is proved. $\square$

Theorem 2.4: *Let feasible set Φ be not empty, condition (16) hold and $L^T Lx \neq 0$ for all $x \in \Phi$. Then problem (6) has a unique solution x^* with positive Lagrange multiplier $\lambda^* > 0$ and $\|Ax^* - b\| = \delta$.*

Proof: The existence of solution x^* follows from Theorem 2.1. The equality $\|Ax^* - b\| = \delta$ and the positiveness of Lagrange multiplier $\lambda^* > 0$ follow from Theorem 2.3. Thus, we just need to prove that x^* is a unique solution.

On the contrary assume that there is another solution y^* that is different from x^* ; $y^* \neq x^*$. Clearly

$$Lx^* = Ly^* = l^* \text{ and } \|Ax^* - b\| = \delta, \|Ay^* - b\| = \delta.$$

Denote $z^* = y^* - x^*$ and consider

$$x_t = x^* + tz^*, \quad t \in [0, 1].$$

Clearly, $\|z^*\| > 0$ and points $\|x_t\|$, $t \in [0, 1]$, belong to the convex combination of edge points x^* and y^* . Since functions $\|Lx\|$ and $\|Ax - b\|$ are convex, we have:

$$\|Lx_t\| \leq l^* \quad \text{and} \quad \|Ax_t - b\| \leq \delta. \quad (18)$$

The last inequality shows that $x_t \in \Phi$ is a feasible point for all $t \in [0, 1]$. Moreover, from (7) we obtain $\|Lx_t\| = l^*$. Then for each $t \in [0, 1]$, x_t is an optimal solution to problem (6) :

$$\|Lx_t\| = l^*, \quad \forall t \in [0, 1]. \quad (19)$$

On the other hand, if $\|Ax_t - b\| < \delta$ for some $t \in [0, 1]$, then gradient of the function $\|Lx\|^2$ at this point should be zero:

$$\nabla\left(\frac{1}{2}\|Lx_t\|^2\right) = L^T Lx_t = 0.$$

This contradicts the assumption of the theorem. Then we have

$$\|Ax_t - b\| = \delta, \quad \forall t \in [0, 1]. \quad (20)$$

Denote

$$\phi(t) \doteq \frac{1}{2}\|Lx_t\|^2, \quad \psi(t) \doteq \frac{1}{2}\|Ax_t - b\|^2.$$

From (19) and (20) we observe that these functions are constant in the interval $t \in [0, 1]$ and therefore their second derivatives are zero:

$$\frac{d^2}{d^2t}\phi(t) = \frac{d^2}{d^2t}\psi(t) \equiv 0, \quad t \in (0, 1). \quad (21)$$

It is not difficult to calculate that

$$\frac{d^2}{d^2t}\phi(t) = (z^*)^T L^T L z^* = \|L z^*\|^2.$$

Then from (21) we obtain $z^* \in \text{Null}(L)$.

Similarly, as

$$\frac{d^2}{d^2t}\psi(t) = (z^*)^T A^T A z^* = \|A z^*\|^2$$

from (21) we obtain $z^* \in \text{Null}(A)$. Therefore,

$$z^* \in \text{Null}(L) \cap \text{Null}(A).$$

Since $z^* = y^* - x^* \neq 0$, this contradicts condition (16).

Theorem is proved. $\square$

It is not difficult to show that if $0 \notin \Phi$ and $\text{rank} L = n$ then all the conditions of Theorem 2.4 are satisfied. Therefore, as a particular case of Theorem 2.4 we have the following result:

Theorem 2.5: *Let $\Phi \neq \emptyset$, $0 \notin \Phi$ and $\text{rank} L = n$. Then problem (6) has a unique solution x^* with positive Lagrange multiplier $\lambda^* > 0$ and $\|Ax^* - b\| = \delta$.*

We note that if $0 \notin \Phi$; that is, if $\|b\| \leq \delta$, then problem (6) has a trivial solution $x^* = 0$.

Under the assumptions of Theorem 2.3, problem (6) is equivalent to the following problem:

$$\text{minimize } \frac{1}{2}\|Lx\|^2, \quad \text{subject to } \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}\delta^2. \quad (22)$$

However in numerical implementations, the problem formulation in the form (6) is more preferable than (22) because of the difficulties related to equality constraints (see section 4).

In this paper, we present a new approach to solving problem (6) for large-scale ill-posed problems. This approach involves the Karush-Kuhn-Tucker conditions and the Fisher-Burmeister function. Similar to the method introduced in [14] and [15], the approach proposed does not require any prior good estimate of the regularization parameter; it is updated iteratively.

Consider problem (6). By the Karush-Kuhn-Tucker conditions, there exists Lagrange multiplier $\lambda \geq 0$, such that

$$\begin{cases} L^T L x + \lambda A^T (Ax - b) = 0, \\ \lambda \geq 0, \quad \delta^2 \geq \|Ax - b\|^2, \quad \lambda(\delta^2 - \|Ax - b\|^2) = 0. \end{cases} \quad (23)$$

We will use the Fisher-Burmeister function [5] $\varphi : R^2 \rightarrow R$ defined by

$$\varphi(a, b) = \sqrt{a^2 + b^2} - a - b.$$

This function satisfies the following property:

$$\varphi(a, b) = 0 \iff a \geq 0, \quad b \geq 0 \quad \text{and} \quad ab = 0.$$

By using this property, we can reformulate system (23) as the following system of nonlinear equations

$$F(z) = 0, \tag{24}$$

where $F : R^{n+1} \rightarrow R^{n+1}$ is defined by

$$F(z) = \begin{pmatrix} L^T Lx + \lambda A^T (Ax - b) \\ \varphi(\lambda, \gamma(\delta^2 - \|Ax - b\|^2)) \end{pmatrix}. \tag{25}$$

Here $\gamma \in R$ is a positive parameter and $z = (x^T, \lambda)^T$. Under the conditions of Theorem 2.3, it follows that the Lagrange multiplier λ corresponding to the solution of the problem (6) is positive: $\lambda > 0$. Then, we obtain that function F is differentiable at the solution of the regularization problem (6). Therefore, the Newton method can naturally be applied to solve nonlinear equation (24)-(25).

The application of the Newton method to solve nonlinear equation (24)-(25) leads to the iterations

$$z_{k+1} = z_k + \Delta z_k, \quad k = 0, 1, 2, \dots, \tag{26}$$

where $z_k = (x_k^T, \lambda_k)^T$, and $\Delta z_k = (\Delta x_k^T, \Delta \lambda_k)^T$ satisfies

$$\begin{pmatrix} L^T L + \lambda_k A^T A & v_k \\ \zeta_k v_k^T & \eta_k \end{pmatrix} \begin{pmatrix} \Delta x_k \\ \Delta \lambda_k \end{pmatrix} = - \begin{pmatrix} L^T Lx_k + \lambda_k A^T (Ax_k - b) \\ \varphi(\lambda_k, \gamma(\delta^2 - \|Ax_k - b\|^2)) \end{pmatrix}. \tag{27}$$

Here we used the following notations:

$$\zeta_k = \frac{-2\gamma^2(\delta^2 - \|Ax_k - b\|^2)}{\sqrt{\lambda_k^2 + \gamma^2(\delta^2 - \|Ax_k - b\|^2)^2}} + 2\gamma, \quad \eta_k = \frac{\lambda_k}{\sqrt{\lambda_k^2 + \gamma^2(\delta^2 - \|Ax_k - b\|^2)^2}} - 1,$$

and $v_k = A^T(Ax_k - b)$.

We have the following

Theorem 2.6: Assume that $\lambda_k > 0$ and condition (16) holds; that is, $Null(L) \cap Null(A) = \emptyset$. Then linear system (27) is well defined.

Proof: It is sufficient to show that the coefficient matrix in system (27) is non-singular.

Since $\lambda_k > 0$ and $Null(L) \cap Null(A) = \emptyset$, matrix $L^T L + \lambda_k A^T A$ is positive definite. Suppose $y_1 \in R^n$ and $y_2 \in R$ such that

$$\begin{pmatrix} L^T L + \lambda_k A^T A & v_k \\ \zeta_k v_k^T & \eta_k \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 0; \tag{28}$$

that is,

$$\begin{cases} (L^T L + \lambda_k A^T A) y_1 + y_2 v_k = 0 \\ \zeta_k v_k^T y_1 + \eta_k y_2 = 0 \end{cases} \quad (29)$$

From the first equation of (29), we have

$$y_1 = -y_2 (L^T L + \lambda_k A^T A)^{-1} v_k. \quad (30)$$

Substituting this in the second equation of (29), we obtain

$$(-\zeta_k v_k^T (L^T L + \lambda_k A^T A)^{-1} v_k + \eta_k) y_2 = 0.$$

Since

$$\zeta_k = \frac{-2\gamma^2(\delta^2 - \|Ax_k - b\|^2)}{\sqrt{\lambda_k^2 + \gamma^2(\delta^2 - \|Ax_k - b\|^2)^2}} + 2\gamma = 2\gamma \left(1 - \frac{\gamma(\delta^2 - \|Ax_k - b\|^2)}{\sqrt{\lambda_k^2 + \gamma^2(\delta^2 - \|Ax_k - b\|^2)^2}}\right) > 0,$$

and

$$\eta_k = \frac{\lambda_k}{\sqrt{\lambda_k^2 + \gamma^2(\delta^2 - \|Ax_k - b\|^2)^2}} - 1 < 0,$$

we have

$$(-\zeta_k v_k^T (L^T L + \lambda_k A^T A)^{-1} v_k + \eta_k) \neq 0.$$

Therefore $y_2 = 0$. From (30), we obtain $y_1 = 0$.

The above discussion shows that, linear system (28) has unique zero solution.

Theorem is proved. $\square$

We note that, in order to globalize the Newton method, a line search technique is used to achieve a sufficient decrease in the natural merit function

$$M(z) = \frac{1}{2} \|F(z)\|^2;$$

where $F(z)$ is defined in (25). Now we describe the main algorithm of this paper.

Algorithm 1.

0. Initialize $z_0 = (x_0^T, \lambda_0)^T \in R^{n+1}$. Set $k = 0$.
1. Computation of the search direction.
Compute $(\Delta x_k, \Delta \lambda_k)$ by solving the following linear equation

$$\begin{pmatrix} L^T L + \lambda_k A^T A & v_k \\ \zeta_k v_k^T & \eta_k \end{pmatrix} \begin{pmatrix} \Delta x_k \\ \Delta \lambda_k \end{pmatrix} = - \begin{pmatrix} L^T L x_k + \lambda_k A^T (Ax_k - b) \\ \varphi(\lambda_k, \gamma(\delta^2 - \|Ax_k - b\|^2)) \end{pmatrix}$$

2. Line search.

ADR Subsets	Number of Features	Number of Data Points	Number of Classes
<i>Cardiovascular</i>	804	21871	5
<i>Blood</i>	1896	8574	4
<i>Body</i>	3518	34456	4
<i>Neuro</i>	3705	50591	4

Table 1. Number of features, data points, and classes of four ADR subsets: Cardiovascular, Blood, Body, and Neuro.

Find the first number α_k of the sequence $\{1, \frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^i}, \dots\}$ satisfying:

- i) $M(x_k + \alpha \Delta x_k, \lambda_k + \alpha_k \Delta \lambda_k) \leq M(x_k, \lambda_k) + \mu \alpha_k (\Delta x_k, \Delta \lambda_k)^T \nabla_{(x, \lambda)} M(x_k, \lambda_k)$ with $\mu = 10^{-4}$
- ii) $\lambda_k + \alpha_k \Delta \lambda_k > 0$

3. Updates.

Set

$$\begin{aligned} x_{k+1} &= x_k + \alpha \Delta x_k, \\ \lambda_{k+1} &= \lambda_k + \alpha_k \Delta \lambda_k. \end{aligned}$$

Set $k = k + 1$ and goto step 1 until one of the following the following stopping criteria is satisfied.

- i) $k \geq k_{max}$
- ii) $\frac{|M(x_{k+1}, \lambda_{k+1}) - M(x_k, \lambda_k)|}{M(x_k, \lambda_k)} \leq \tau_1$
- iii) $\|\alpha_k \begin{pmatrix} \Delta x_k \\ \Delta \lambda_k \end{pmatrix}\| \leq \tau_2$
- iv) $\|Ax_k - b\|^2 \leq \delta^2$

It is clear that if $rank L = n$ then condition (16) holds. In this case from Theorems 2.5 and 2.6 we obtain that **Algorithm 1** can be applied without making any assumptions with respect to matrix A .

3. Adverse Drug Reaction (ADR) Problems

The Australian Adverse Drug Reaction Advisory Committee (ADRAC) database developed and maintained by Therapeutic Goods Administration (TGA) is aimed at the early detection of adverse drug reaction signals. Records of prescribed drugs used on patients from 1972 to 2001 were accumulated in the database. In the numerical experiments we will consider four different sets of reactions called *Cardiovascular*, *Blood*, *Body*, and *Neuro* (see [17] for more details).

Drugs taken by patients are considered as *features* and reactions caused by these drugs are considered as *classes*. Therefore, matrix A in the regularization problems has elements a_{ij} , where $a_{ij} = 1$ if drug j is used by patient i , and $a_{ij} = 0$ if not.

From Figure 1 we can see that these four ADR subsets are large scale data sets. Each record can have more than one adverse reactions (classes), and at the same time, the number of drugs taken by patients is highly limited: almost 50 percent of patients use only one feature (drug). Therefore, this type of data sets can be considered as examples for SFML data sets [18].

The problem of finding the relationship between drugs and adverse drug reactions is one of the challenging problems in ADR. In this paper, these relationships will be described by a solution matrix X to problem (6). This approach leads to a linear model when considering the adverse reaction effects from a set of drugs. In [17] it is shown that such a linear model is well suitable for the study of many ADR problems including drug-drug interactions.

The methods developed in our previous studies (see [17] and references therein) requires that matrix X has non-negative elements only. This means that only positive adverse effects from drugs can be considered by these methods. In reality, some drugs may have “negative” (“healing”) effects as well. This requires considering matrices X with real valued elements that could be obtained, for instance, from the linear least-squares formulations.

N	<i>Cardiovascular</i>	<i>Blood</i>	<i>Body</i>	<i>Neuro</i>
10	1.73	1.73	2.45	2.36
20	3.73	3.23	4.73	2.57
30	4.79	4.13	5.45	4.29
40	5.36	6.88	5.95	6.23
50	9.49	7.56E+15	6.37	6.65
60	9.49	9.78E+15	8.61	9.42
70	10.9	9.91E+15	9.96	10.3
80	14	2.41E+16	1.52E+16	1.67E+16
90	17.5	9.03E+15	3.73E+16	1.99E+16
100	18.6	1.92E+16	1.82E+16	2.01E+16
200	1.87E+17	9.61E+16	Inf	1.80E+17

Table 2. The condition number $Cond(A)$ of matrix A corresponding to the first N records: $N = 10, 20, \dots$

Table 2 presents the condition number of matrix A for different number of records after deleting linear independent rows. We can observe that starting from the first 200 records, the condition number of corresponding matrix A in all data sets becomes very large; that is, the corresponding optimization problems become ill-posed.

Therefore, the linear least-squares lead to highly ill-posed problems. This is the main reason that such models have not been applied to the study of ADR problems yet. In this paper we consider a regularized formulation in the form (6). A solution to this problem provides drug-reaction relationships X that may have negative elements.

4. Numerical Experiments

All the numerical experiments run in MATLAB. We used the identity matrix for L in (6) when applying Algorithm 1 introduced above. The Newton system (27) is solved by the Generalized Minimal Residual method (GMRES) from MATLAB Toolbox “gmres.m” with the default tolerance 10^{-6} . This system is well-conditioned by Theorem 2.6. Together with this algorithm, we also apply the following two methods for comparison.

Conjugate Gradients Least Squares (CGLS) proposed by Hestenes and Stiefel [10] is a well-known method to the least squares problem, especially for solving sparse systems of equations with a symmetric positive definite coefficient matrix. The conjugate gradient method is iterative in nature, so it can be applied to sparse systems that are too large to be handled by direct methods such as Tikhonov

regularization.

We also apply the hybrid method with W-GCV regularization (denoted as HyBR) proposed recently in [3]. It is proven to be a very efficient method for solving large-scale ill-posed image restoration problems.

The parameters are chosen as follows. For Algorithm 1, we set x_0 to be the zero vector, $\lambda_0 = 1$, $\tau_1 = 10^{-4}$ and $\tau_2 = 10^{-8}$. All the algorithms are stopped after 50 iterations. HyBR method is stopped when (see [3] for more details)

$$\frac{|GCV(k) - GCV(k-1)|}{GCV(1)} \leq 10^{-6} \text{ or } k > 50.$$

We apply 4-fold cross validation. All the results presented in this section are averaged over these 4 folds. To investigate the effect of discrepancy, different factors $\delta = k\delta'$ for Algorithm 1 are considered, where $k=0.95, 1.00, 1.05, 1.10, 1.15, 1.20$, and δ' is the corresponding values obtained by CGLS (they are almost the same as those for HyBR).

Accuracy of the classification is measured by the *Average Precision* [16, 22] that is a performance measure frequently used in Information Retrieval [23]. In our experiments, we use a modified version of the Average Precision [17] described below to evaluate the performance of different classifiers.

4.1. Evaluation measure: Average Precision

Given a data point (sample) $a = (a_1, \dots, a_n)$, denote by $\mathcal{Y}(a)$ an c -dimensional vector of classes with elements 0 and 1, and let $Y(a) = \{l \in \{1, \dots, c\} : \mathcal{Y}_l(a) = 1\}$ be the set of actual classes.

Consider a classifier $\mathcal{H}(a)$ that defines a prediction vector

$$\mathcal{H}(a) = (\mathcal{H}_1(a), \dots, \mathcal{H}_c(a))$$

with real valued components $\mathcal{H}_l(a)$, $l = 1, \dots, c$. Higher values of the components correspond to higher relevance to the classes in the prediction.

If X is a solution to (6), then classifier $\mathcal{H}(a)$ is defined as $\mathcal{H}(a) = a^T X$.

We denote by $\mathcal{T}(a)$ the set of all ordered classes $\tau = \{i_1, \dots, i_c\}$ satisfying the condition

$$\mathcal{H}_{i_1}(a) \geq \dots \geq \mathcal{H}_{i_c}(a);$$

where $i_k \in \{1, \dots, c\}$ and $i_k \neq i_m$ if $k \neq m$.

In the case, when numbers $\mathcal{H}_i(a)$, $i = 1, \dots, c$, are different, there is just one order τ satisfying this condition. But if there are classes having the same value then we can order the classes in different ways; that is, in this case the set $\mathcal{T}(a)$ contains more than one order.

Given order $\tau = \{\tau_1, \dots, \tau_c\} \in \mathcal{T}(a)$, we define the rank for each class $l \in Y(a)$ as $rank_\tau(a; l) = k$, where the number k satisfies $\tau_k = l$. Then *Precision* is defined as:

$$P_\tau(a) = \frac{1}{|Y(a)|} \sum_{l \in Y(a)} \frac{|\{k \in Y(a) : rank_\tau(a; k) \leq rank_\tau(a; l)\}|}{rank_\tau(a; l)}.$$

Here, we use the notation $|S|$ for the cardinality of the set S . This measure has

the following meaning. For instance, if all observed classes $Y(a)$ have occurred on the top of ordering τ then $P_\tau(a) = 1$. Clearly the number $P_\tau(a)$ depends on order τ . We define

$$P_{best}(a) = \max_{\tau \in T(a)} P_\tau(a) \quad \text{and} \quad P_{worst}(a) = \min_{\tau \in T(a)} P_\tau(a),$$

which are related to the “best” and “worst” ordering. Therefore, it is sensible to define the *Precision* as the midpoint of these two versions:

$$P(a) = \frac{P_{best}(a) + P_{worst}(a)}{2}.$$

Average Precision over a given set of records $\mathcal{A}$ will be defined as:

$$P_{av} = \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{X}} P(a).$$

4.2. The choice of γ

Numerical experiments with different values for γ show that the choice of an appropriate values plays an important role. This is mainly related to the convergence rate of the algorithm around the “true” optimal solution when dealing with large-scale ill-posed problems. This situation is typical (for example in image restoration problems) in the application of regularization methods and it requires using very large values for γ . Therefore, a rational value of γ in the application to ADR datasets might also be important.

γ	#iterations	$\delta^2 - \ Ax - b\ ^2$	$\ x\ ^2$
10^3	100	0.0605	14.2401
10^4	100	0.0533	14.3724
10^5	100	0.0446	14.3774
10^6	10	0.0383	14.4393
10^7	8	0.1604	13.1485
10^8	30	0.0000	12.9340
10^9	43	0.0000	13.2184
10^{10}	31	0.0230	12.7203

Table 3. Results obtained for different values of γ on *Body* data set. In the algorithm the stopping criteria is set as follows: $k_{max} = 100$ and the last criteria (iv) is removed.

Table 3 shows the effect of γ in the *Body* data set. For the sake of simplicity, we provide the results for only one class out of four classes. Stopping criteria is defined as follows: we set the maximum number of iteration as $k_{max} = 100$ and remove the last criteria (iv) $\|Ax_k - b\| \leq \delta$. The results obtained for different numbers γ are presented in Table 3.

We observe that, when γ is small, for example $10^3, 10^4$ and 10^5 , the algorithm stops reaching the maximal number for iterations (i.e., 100). After γ is increased, the algorithm stops at a “reasonable” number of iterations achieving one of the stopping criteria (ii) and (iii). The best results is achieved at $\gamma = 10^{10}$ in the sense that the value of $\|x\|^2$ is minimal. Taking this into account, in all experiment below, we set $\gamma = 10^{10}$ that turns out to be a suitable choice for other datasets as well.

<i>Cardiovascular</i>	$\ AX - B\ $	$\ X\ $	$\ AX - B\ + \ X\ $	Average Precision
CGLS	121.50	49.73	171.88	76.95
Hybr	121.89	34.71	156.60	77.19
Algorithm 1 ($k = 0.95$)	115.57	49.73	165.31	77.87
Algorithm 1 ($k = 1.00$)	116.89	36.15	153.04	77.09
Algorithm 1 ($k = 1.05$)	122.51	20.55	143.06	78.07
Algorithm 1 ($k = 1.10$)	128.35	13.14	141.49	78.64
Algorithm 1 ($k = 1.15$)	134.14	9.42	143.56	78.55
Algorithm 1 ($k = 1.20$)	139.83	8.29	148.11	77.97

Table 4. *Cardiovascular* dataset: Regularization factors for Algorithm 1 for different values $\delta = k\delta'$, where δ' is the corresponding values obtained by CGLS (they are almost the same as those for HyBR). The values presented are averaged over the 4 folds.

An interesting observation is that inside the feasible region $\|Ax_k - b\| \leq \delta$ the stopping criteria (ii) or (iii) can be easily reached which indicates very small improvements in next iterations. Continuing with such small improvements is time consuming, especially for large-scale ill-posed problems. Taking into account this issue, it is quite common in the literature to apply an additional stopping criteria $\|Ax_k - b\| \leq \delta$. We will also use this criteria (i.e., (iv)) in all our experiments below.

4.3. Classification with ADRAC data

The experimental results for the four ADRAC subsets are listed in Table 4-7. Four factors are considered here including data fitting factor $\|AX - B\|$, regularization factor $\|X\|$, the summation of these two factors, and also classification accuracies in terms of the *Average Precision*. We provide the values of sum $\|AX - B\| + \|X\|$ in order to show the tradeoff of data fitting factor $\|AX - B\|$ and regularization factor $\|X\|$. For Algorithm 1 we considered different δ values by taking $\delta = k\delta'$, $k = 0.95, 1.00, 1.05, 1.10, 1.15, 1.20$, where δ' is the value of $\|AX - B\|$ obtained by *CGLS*.

We observe that *CGLS* performs better in fitting the training data in terms of the values $\|AX - B\|$ that are generally smaller compared with *Hybr* and Algorithm 1. However, the function smoothness factor $\|X\|$ of *CGLS* is the worst which means that the minimizer calculated by *CGLS* is not well regularized. As a result, *CGLS* archives lower classification accuracies on test sets for the all the four ADR data sets.

We can also observe an expected pattern from Tables 4-7 that the relationship between $\|AX - B\|$ and $\|X\|$ are negative. Well balanced classifiers between data fitting and regularization can have higher performances in the classification of new examples.

We observe that, compared to CGLS and HyBR, Algorithm 1 archives higher classification accuracies on new examples for all the four data sets when data fitting has a higher level of relaxation; that is, when $\delta = k\delta'$ with $k \geq 1.05$. On the other hand, as it could be expected, when the relaxation in data fitting is not "enough" (that is, $k \leq 1$) the performers of Algorithm 1 decreases in predicting new examples.

Therefore, the results in Tables 4-7 show the importance of applying regularization methods in the establishment of better drug-reaction relationships expressed as a matrix X .

<i>Blood</i>	$\ AX - B\ $	$\ X\ $	$\ AX - B\ + \ X\ $	Average Precision
CGLS	56.09	33.71	89.80	77.03
Hybr	56.99	23.02	80.00	77.70
Algorithm 1 ($k = 0.95$)	56.01	34.14	90.15	76.94
Algorithm 1 ($k = 1.00$)	56.14	32.77	88.92	77.10
Algorithm 1 ($k = 1.05$)	58.66	17.39	76.05	78.10
Algorithm 1 ($k = 1.10$)	61.45	11.86	73.30	78.75
Algorithm 1 ($k = 1.15$)	64.15	8.20	72.45	79.29
Algorithm 1 ($k = 1.20$)	67.02	5.98	73.00	79.56

Table 5. *Blood* dataset: Regularization factors for Algorithm 1 for different values $\delta = k\delta'$, where δ' is the corresponding values obtained by CGLS (they are almost the same as those for HyBR). The values presented are averaged over the 4 folds.

<i>Body</i>	$\ AX - B\ $	$\ X\ $	$\ AX - B\ + \ X\ $	Average Precision
CGLS	111.11	30.20	141.31	90.85
Hybr	112.46	27.55	140.01	91.14
Algorithm 1 ($k = 0.95$)	109.07	43.77	152.84	90.73
Algorithm 1 ($k = 1.00$)	110.29	32.15	142.44	90.94
Algorithm 1 ($k = 1.05$)	115.78	19.97	135.75	91.73
Algorithm 1 ($k = 1.10$)	121.29	14.51	135.80	91.83
Algorithm 1 ($k = 1.15$)	126.61	11.81	138.42	91.47
Algorithm 1 ($k = 1.20$)	131.87	11.05	142.92	91.31

Table 6. *Body* dataset: Regularization factors for Algorithm 1 for different values $\delta = k\delta'$, where δ' is the corresponding values obtained by CGLS (they are almost the same as those for HyBR). The values presented are averaged over the 4 folds.

<i>Neuro</i>	$\ AX - B\ $	$\ X\ $	$\ AX - B\ + \ X\ $	Average Precision
CGLS	171.46	32.09	203.54	73.51
Hybr	170.58	46.55	217.13	73.42
Algorithm 1 ($k = 0.95$)	170.60	45.49	216.10	73.43
Algorithm 1 ($k = 1.00$)	171.51	32.73	204.24	73.40
Algorithm 1 ($k = 1.05$)	180.09	11.22	191.31	73.79
Algorithm 1 ($k = 1.10$)	188.55	6.00	194.56	74.42
Algorithm 1 ($k = 1.15$)	196.99	3.68	200.66	74.17
Algorithm 1 ($k = 1.20$)	205.10	4.16	209.26	74.11

Table 7. *Neuro* dataset: Regularization factors for Algorithm 1 for different values $\delta = k\delta'$, where δ' is the corresponding values obtained by CGLS (they are almost the same as those for HyBR). The values presented are averaged over the 4 folds.

5. Conclusions

The linear least squares fitting (LLSF) technique and its regularized implementations are commonly used in many areas including different problems in Finance like ricing options and eco-finance and gene-environment networks. In this paper a new numerical method for solving regularization with discrepancy is presented. It is based on the Fisher-Burmeister function and the Karush-Kuhn-Tucker conditions, where The Newton method is applied for solving corresponding systems of equations. Numerical results obtained on adverse drug reaction problems demonstrate the efficiency of the proposed approach. The algorithm introduced is more accurate in classification of new examples compared to the methods that aim to achieve better data fitting.

Similar mathematical formulations involving regularization methods with discrepancy principle have been considered in the literature recently in image restora-

tion problems. In this paper we apply our method to adverse drug reaction problems that outline a new class of problems requiring advanced data fitting techniques. The applications of these techniques to financial problems would be an interesting topic for future investigations.

References

- [1] A. Björck. Numerical methods for least squares problems. Society for Industrial Mathematics, 1996.
- [2] A. R. Choudhury, A. King, S. Kumar and Y. Sabharwal. Optimizations in financial engineering: The Least-Squares Monte Carlo method of Longstaff and Schwartz. *In: Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium*, 2008, 1-11.
- [3] J. Chung, J.G. Nagy and D.P. O’Leary. A weighted gcv method for lanczos-hybrid regularization. *Electronic Transactions on Numerical Analysis*, 28:149–167, 2008.
- [4] L. Eldén. Algorithms for the regularization of ill-conditioned least squares problems. *BIT*, 17:134–145, 1977.
- [5] A. Fischer. A special newton-type optimization method. *Optimization*, 24:269–284, 1992.
- [6] J. Hadamard. *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. Dover, New York, 1952.
- [7] P.C. Hansen. Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1–35, 1994.
- [8] P.C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. Polyteknisk Forlag, Lyngby, Denmark, 1996. Doctoral Dissertation.
- [9] P.C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34: 561–580, December 1992.
- [10] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.
- [11] A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.
- [12] V.V. Ivanov. *The theory of approximate methods and their application to the numerical solution of singular integral equations*. Springer, 1976. ISBN 9028600361.
- [13] E. Kropat, G.W. Weber, B.A. Ozturk. Eco-Finance Networks under Uncertainty. *In Proc. EngOpt 2008, The International Conference on Engineering Optimization, Rio de Janeiro, Brazil, June 01-05*, 2008.
- [14] G. Landi. Lagrangian methods for the regularization of discrete ill-posed problems. *Computational Optimization and Applications*, 8:347–368, 2008.
- [15] G. Landi and E.L. Piccolomini. An iterative Lagrange method for the regularization of discrete ill-posed inverse problems. *Computers and Mathematics with Applications*, 60, 1723–1738, 2010.
- [16] T. Li, C. Zhang and S. Zhu. Empirical studies on multi-label classification. *In ICTAI*, pages 86–92. IEEE Computer Society, 2006.
- [17] MA Mammadov, AM Rubinov, and J. Yearwood. The study of drug-reaction relationships using global optimization techniques. *Optimization Methods and Software*, 22(1):99–126, 2007.
- [18] M. Mammadov, J. Yearwood, and A. Banarjee. Classification on shorter featured and multi-label datasets. *In Proc. ICOTA7, The 7-th International Conference on Optimization: Techniques and Applications, Kobe, Japan, December 12-15*, 2007.
- [19] V.A. Morozov, Z. Nashed, and A.B. Aries. *Methods for solving incorrectly posed problems*. Springer New York, 1984. ISBN 3540960597.
- [20] A. Ozmen, G.W. Weber and I. Batmaz. The new robust CMARS (RCMARS) method. *In Proceedings of 24th MEC-EuroPT 2010-Continuous Optimization and Information-Based Technologies in the Financial Sector*, Izmir, Turkey, 2010, pp. 362-368.
- [21] J.E. Pierce and B.W. Rust. Constrained least squares interval estimation. *SIAM Journal on Scientific and Statistical Computing*, 6(3):670–683, July 1985.
- [22] R.E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.
- [23] G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):974–980, 1991.
- [24] X. Tian and K. Benkrid. American option pricing on reconfigurable hardware using least-squares Monte Carlo method. *In Proc. Int. Conf. on Field-Programmable Technology*, p.263-270, 2009.
- [25] A.N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39(5):195–198, 1943.
- [26] A.H. Tse, D.B. Thomas, K. Tsoi, and W. Luk. Efficient reconfigurable design for pricing Asian options. *in HEART’10*, 2010.
- [27] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *ICML*, pages 412–420. Morgan Kaufmann, 1997.