

Chapter 2

Least Squares Problems

Abstract In this chapter we provide an overview of the original minimum least squares problem and its variations. We present their robust formulations as they have been proposed in the literature so far. We show the analytical solutions for each variation and we conclude the chapter with some numerical techniques for computing them efficiently.

2.1 Original Problem

In the original linear least squares (LLS) problem one needs to determine a linear model that approximates “best” a group of samples (data points). Each sample might correspond to a group of experimental parameters or measurements and each individual parameter to a feature or, in statistical terminology, to a predictor. In addition, each sample is characterized by an outcome which is defined by a real valued variable and might correspond to an experimental outcome. Ultimately we wish to determine a linear model able to issue outcome prediction for new samples. The quality of such a model can be determined by a minimum distance criterion between the samples and the linear model. Therefore if n data points, of dimension m each, are represented by a matrix $A \in \mathbb{R}^{n \times m}$ and the outcome variable by a vector $b \in \mathbb{R}^n$ (each entry corresponding to a row of matrix A), we need to determine a vector $x \in \mathbb{R}^m$ such that the residual error, expressed by some norm, is minimized. This can be stated as:

$$\min_x \|Ax - b\|_2^2 \quad (2.1)$$

where $\|\cdot\|_2$ is the Euclidean norm of a vector. The objective function value is also called residual and denoted $r(A, b, x)$ or just r . The geometric interpretation of this problem is to find a vector x such that the sum of the distances between the points represented by the rows of matrix A and the hyperplane defined by $x^T w - b = 0$ (where w is the independent variable) is minimized. In this sense this problem is a

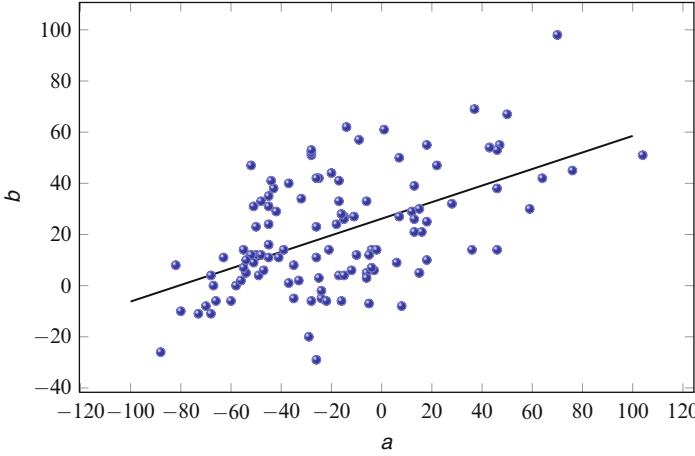


Fig. 2.1 The single input single outcome case. This is a 2D example the predictor represented by the a variable and the outcome by vertical axis b

first order polynomial fitting problem. Then by determining the optimal x vector will be able to issue predictions for new samples by just computing their inner product with x . An example in two dimensions (2D) can be seen in Fig. 2.1. In this case the data matrix will be $A = [a \ e] \in \mathbb{R}^{n \times 2}$ where a is the predictor variable and e a column vector of ones that accounts for the constant term.

The problem can be solved, in its general form, analytically since we know that the global minimum will be at a Karush–Kuhn–Tucker (KKT) point (since the problem is convex and unconstrained) the Lagrangian equation $\mathcal{L}_{\text{LLS}}(x)$ will be given by the objective function itself and the KKT points can be obtained by solving the following equation:

$$\frac{d\mathcal{L}_{\text{LLS}}(x)}{dx} = 0 \Leftrightarrow 2A^T Ax = A^T b \quad (2.2)$$

In case that A is of full row rank, that is $\text{rank}(A) = n$, matrix $A^T A$ is invertible and we can write:

$$x_{\text{LLS}} = (A^T A)^{-1} A^T b \triangleq A^\dagger b \quad (2.3)$$

Matrix A^\dagger is also called pseudoinverse or Moore–Penrose matrix. It is very common that the full rank assumption is not always valid. In such case the most common way to address the problem is through regularization. One of the most famous regularization techniques is the one known as Tikhonov regularization [55]. In this case instead of problem (2.1) we consider the following problem:

$$\min_x (\|Ax - b\|^2 + \delta \|x\|^2) \quad (2.4)$$

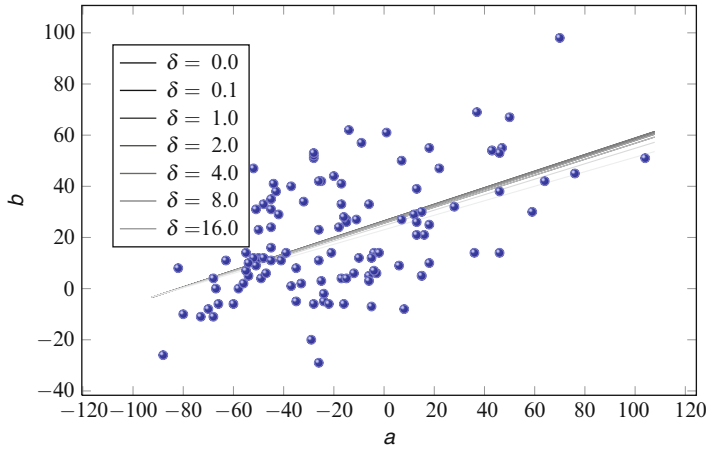


Fig. 2.2 LLS and regularization. Change of linear least squares solution with respect to different δ values. As we can observe, in this particular example, the solution hyperplane is slightly perturbed for different values of δ

by using the same methodology we obtain:

$$\frac{d\mathcal{L}_{\text{RLLS}}(x)}{dx} = 0 \Leftrightarrow A^T(Ax - b) + \delta Ix = 0 \Leftrightarrow (A^T A + \delta I)x = A^T b \quad (2.5)$$

where I is a unit matrix of appropriate dimension. Now even in case that $A^T A$ is not invertible we can compute x by

$$x_{\text{RLLS}} = (A^T A + \delta I)^{-1} A^T b \quad (2.6)$$

This type of least square solution is also known as ridge regression. The parameter δ controls the trade-off between optimality and stability. Originally regularization was proposed in order to overcome this practical difficulty that arises in real problems and it is related to rank deficiency described earlier. The value of δ is determined usually by trial and error and its magnitude is smaller compared to the entries of data matrix. In Fig. 2.2 we can see how the least squares plane changes for different values of delta.

In Sect. 2.5 we will examine the relation between robust linear least squares and robust optimization.

2.2 Weighted Linear Least Squares

A slight, and more general, modification of the original least squares problem is the weighted linear least squares problem (WLLS). In this case we have the following minimization problem:

$$\min_x r^T W r = \min_x (Ax - b)^T W (Ax - b) = \min_x \|W^{1/2}(Ax - b)\| \quad (2.7)$$

where W is the weight matrix. Note that this is a more general formulation since for $W = I$ the problem reduces to (2.1). The minimum can be again obtained by the solution of the corresponding KKT systems which is:

$$2A^T W (Ax - b) = 0 \quad (2.8)$$

and gives the following solution:

$$x_{\text{WLLS}} = (A^T W A)^{-1} A^T W b \quad (2.9)$$

assuming that $A^T W A$ is invertible. If this is not the case regularization is employed resulting in the following regularized weighted linear least squares (RWLLS) problem

$$\min_x \left(\|W^{1/2}(Ax - b)\|^2 + \delta \|x\|^2 \right) \quad (2.10)$$

that attains its global minimum for

$$x_{\text{RWLLS}} = (A^T W A + \delta I)^{-1} A^T W b \quad (2.11)$$

Next we will discuss some practical approaches for computing least square solution for all the discussed variations of the problem.

2.3 Computational Aspects of Linear Least Squares

Least squares solution can be obtained by computing an inverse matrix and applying a couple of matrix multiplications. However, in practice, direct matrix inversion is avoided, especially due to the high computational cost and solution instabilities. Here we will describe three of the most popular methods used for solving the least squares problems.

2.3.1 Cholesky Factorization

When matrix A is of full rank, then AA^T is invertible and can be decomposed through Cholesky decomposition in a product LL^T where L is a lower triangular matrix. Then (2.2) can be written as:

$$LL^T x = A^T b \quad (2.12)$$

that can be solved by a forward substitution followed by a backward substitution. In case that A is not of full rank, then this procedure can be applied to the regularized problem (2.5).

2.3.2 QR Factorization

An alternative method is the one of QR decomposition. In this case we decompose matrix AA^T into a product of two matrices where the first matrix Q is orthogonal and the second matrix R is upper triangular. This decomposition again requires data matrix A to be of full row rank. Orthogonal matrix Q has the property $QQ^T = I$ thus the problem is equivalent to

$$Rx = Q^T A^T b \quad (2.13)$$

and it can be solved by backward substitution.

2.3.3 Singular Value Decomposition

This last method does not require full rank of matrix A . It uses the singular value decomposition of A :

$$A = U\Sigma V^T \quad (2.14)$$

where U and V are orthogonal matrices and Σ is diagonal matrix that has the singular values. Every matrix with real elements has an SVD and furthermore it can be proved that a matrix is of full row rank if and only if all of its singular values are nonzero. Substituting with its SVD decomposition we get:

$$AA^T x = (U\Sigma V^T)(V\Sigma U^T)x = U\Sigma^2 U^T x = A^T b \quad (2.15)$$

and finally

$$x = U(\Sigma^2)^\dagger U^T A^T b \quad (2.16)$$

The matrix $(\Sigma^2)^\dagger$ can be computed easily by inverting its nonzero entries. If A is of full rank then all singular values are non-zero and $(\Sigma^2)^\dagger = (\Sigma^2)^{-1}$. Although SVD can be applied to any kind of matrix it is computationally expensive and sometimes is not preferred especially when processing massive datasets.

2.4 Least Absolute Shrinkage and Selection Operator

An alternative regularization technique for the same problem is the one of least absolute shrinkage and selection operator (LASSO) [54]. In this case the regularization term contains a first norm term $\delta\|x\|_1$. Thus we have the following minimization problem:

$$\min_x (\|Ax - b\|^2 + \delta\|x\|_1) \quad (2.17)$$

Although this problem cannot be solved analytically as the one obtained after Tikhonov regularization, sometimes it is preferred as it provides sparse solutions. That is the solution x vector obtained by LASSO has more zero entries. This approach has a lot of applications in compressive sensing [2, 16, 34]. As it will be discussed later this regularization possesses further robust properties as it can be obtained through robust optimization for a specific type of data perturbations.

2.5 Robust Least Squares

2.5.1 Coupled Uncertainty

Now we will study the robust version of problem (2.1). The results presented here were first described in [20] and similar results were independently obtained in [18]. At the end we describe some extensions that were first described in [17]. As we discussed earlier the RC formulation of a problem involves solution of a worst case scenario problem. This is expressed by a min–max (or max–min) type problem where the outer min (max) problem refers to the original one whereas the inner max (min) to the worst admissible scenario. For the least squares case the generic RC formulation can be described from the following problem:

$$\min_x \max_{\Delta A \in \mathcal{U}_A, \Delta b \in \mathcal{U}_b} \|(A + \Delta A)x - (b + \Delta b)\|_2 \quad (2.18)$$

where $\Delta A, \Delta b$ are perturbation matrices and $\mathcal{U}_A, \mathcal{U}_b$ are sets of admissible perturbations. As in many robust optimization problems, the structural properties of $\mathcal{U}_A, \mathcal{U}_b$ are important for the computational tractability of the problem. Here we study the case where the two perturbation matrices are unknown but their norm is bounded by a known constant. Thus we have the following optimization problem:

$$\min_x \max_{\|\Delta A\| \leq \rho_A, \|\Delta b\| \leq \rho_b} \|(A + \Delta A)x - (b + \Delta b)\|_2 \quad (2.19)$$

This type of uncertainty is often called coupled uncertainty because the uncertainty information is not given in terms of each sample individually but in terms of the whole data matrix. This can be interpreted as having a total uncertainty

“budget” which not required to be distributed evenly among the dataset. Under this assumption we do not have any particular information for individual data points and the resulting solution to this problem can be extremely conservative. First we will reduce problem (2.19) to a minimization problem through the following lemma

Lemma 2.1. *The problem (2.19) is equivalent to the following:*

$$\min_x (\|Ax - b\| + \rho_A \|x\| + \rho_b) \quad (2.20)$$

Proof. From triangular inequality we can obtain an upper bound on the objective function of (2.19):

$$\|(A + \Delta A)x - (b + \Delta b)\| \leq \|Ax - b\| + \|\Delta Ax - \Delta b\| \quad (2.21)$$

$$\leq \|Ax - b\| + \|\Delta A\| \|x\| + \|\Delta b\| \quad (2.22)$$

$$\leq \|Ax - b\| + \rho_A \|x\| + \rho_B \quad (2.23)$$

Now if in the original problem (2.19) we set

$$\Delta A = \frac{Ax - b}{\|Ax - b\|} \frac{x^T}{\|x\|} \rho_A, \quad \Delta b = -\frac{Ax - b}{\|Ax - b\|} \rho_B \quad (2.24)$$

we get

$$\begin{aligned} \|(A + \Delta A)x - (b + \Delta b)\| &= \|Ax - b + \Delta Ax - \Delta b\| \\ &= \|Ax - b\| \left(1 + \frac{\|x\|}{\|Ax - b\|} \rho_A + \frac{1}{\|Ax - b\|} \rho_B \right) \\ &= \|Ax - b\| + \rho_A \|x\| + \rho_B \end{aligned} \quad (2.25)$$

This means that the upper bound obtained by the triangular inequality can be achieved by (2.24). Since the problem is convex, this will be its global optimum.

We can easily observe that the point (2.24) satisfies the optimality conditions. Since problem (2.20) is unconstrained, its Lagrangian will be the same as the cost function. Since this function is convex we just need to examine the points for which the derivative is equal to zero and consider separate cases for the non-differentiable points. At the points where the cost function is differentiable we have:

$$\frac{\partial \mathcal{L}_{\text{RLLS}}(x)}{\partial x} = 0 \Leftrightarrow \frac{A^T(Ax - b)}{\|Ax - b\|} + \frac{x}{\|x\|} \rho_A = 0 \quad (2.26)$$

From this last expression we require $x \neq 0$ and $Ax \neq b$ (we will deal with this cases later). If we solve with respect to x , we obtain:

$$\frac{1}{\|Ax - b\|} \left(A^T(Ax - b) + x \frac{\|Ax - b\|}{\|x\|} \rho_A \right) = 0 \quad (2.27)$$

or

$$\left(A^T A + \rho_A \frac{\|Ax - b\|}{\|x\|} I \right) x = A^T b \quad (2.28)$$

and finally

$$x = (A^T A + \mu I)^{-1} A^T b, \text{ where } \mu = \frac{\|Ax - b\|}{\|x\|} \rho_A \quad (2.29)$$

In case that $Ax = b$ the solution is given by $x = A^\dagger b$ where A^\dagger is the Moore–Penrose or pseudoinverse matrix of A . Therefore we can summarize this result in the following lemma:

Lemma 2.2. *The optimal solution to problem (2.20) is given by:*

$$x = \begin{cases} A^\dagger b & \text{if } Ax = b \\ (A^T A + \mu I)^{-1} A^T b, \mu = \rho_A \frac{\|Ax - b\|}{\|x\|} & \text{otherwise} \end{cases} \quad (2.30)$$

Since in this last expression μ is a function of x we need to provide a way in order to tune it. For this we need to use the singular value decomposition of data matrix A :

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T \quad (2.31)$$

where Σ is the diagonal matrix that contains the singular values of A in descending order. In addition we partition the vector $U^T b$ as follows:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = U^T b \quad (2.32)$$

where b_1 contains the first n elements and b_2 the rest $m - n$. Now using this decompositions we will obtain two expressions for the numerator and the denominator of μ . First for the denominator:

$$x = (A^T A + \mu I)^{-1} A^T b = (V \Sigma^2 V^T + \mu I)^{-1} V \Sigma b_1 = V (\Sigma^2 + \mu I)^{-1} \Sigma b_1 \quad (2.33)$$

the norm will be given from

$$\|x\| = \|\Sigma (\Sigma^2 + \mu I)^{-1}\| \quad (2.34)$$

and for the numerator

$$Ax - b = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T V (\Sigma^2 + \mu I)^{-1} \Sigma b_1 - b \quad (2.35)$$

$$= U \left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix} (\Sigma^2 + \mu I)^{-1} \Sigma b_1 - U^T b \right) \quad (2.36)$$

$$= U \left(\begin{bmatrix} \Sigma(\Sigma^2 + \mu I)^{-1} \Sigma b_1 - b_1 \\ -b_2 \end{bmatrix} \right) \quad (2.37)$$

$$= U \begin{bmatrix} -\mu(\Sigma^2 + \mu I)^{-1} b_1 \\ -b_2 \end{bmatrix} \quad (2.38)$$

and for the norm

$$\|Ax - b\| = \sqrt{\|b_2\|^2 + \alpha^2 \|(\Sigma^2 + \mu I)^{-1} b_1\|^2} \quad (2.39)$$

Thus μ will be given by:

$$\mu = \frac{\|Ax - b\|}{\|x\|} = \rho_A \frac{\sqrt{\|b_2\|^2 + \alpha^2 \|(\Sigma^2 + \mu I)^{-1} b_1\|^2}}{\|\Sigma(\Sigma^2 + \mu I)^{-1} b_1\|} \quad (2.40)$$

Note that in the present analysis we assume that data matrix A is of full rank. If this is not the case similar analysis can be performed (for details, see [17]). The final solution can be obtained by the solution of (2.40) computationally. Next we will present some variations of the original least squares problem that are discussed in [17].

2.6 Variations of the Original Problem

In [17] authors introduced least square formulation for slightly different perturbation scenarios. For example, in the case of the weighted least squares problem with weight uncertainty one is interested to find:

$$\min_x \max_{\|\Delta W\| \leq \rho_W} \|(W + \Delta W)(Ax - b)\| \quad (2.41)$$

using the triangular inequality we can obtain an upper bound:

$$\|(W + \Delta W)(Ax - b)\| \leq \|W(Ax - b)\| + \|\Delta W(Ax - b)\| \quad (2.42)$$

$$\leq \|W(Ax - b)\| + \rho_W \|Ax - b\| \quad (2.43)$$

Thus the inner maximization problem reduces to the following problem:

$$\min_x (\|W(Ax - b)\| + \rho_W \|Ax - b\|) \quad (2.44)$$

by taking the corresponding KKT conditions, similar to previous analysis, we obtain:

$$\frac{\partial \mathcal{L}_{\text{WLLS}}(x)}{\partial x} = \frac{\partial \|W(Ax - b)\|}{\partial x} + \frac{\partial \|Ax - b\|}{\partial x} \quad (2.45)$$

$$= \frac{A^T W^T (WAx - Wb)}{\|W(Ax - b)\|} + \rho_W \frac{A^T (Ax - b)}{\|Ax - b\|} \quad (2.46)$$

By solving the equation

$$\frac{\partial \mathcal{L}_{\text{WLLS}}(x)}{\partial x} = 0 \quad (2.47)$$

we find that the solution should satisfy

$$A^T (W^T W + \mu I) Ax = A^T (W^T W + \mu I) b \quad \text{where} \quad \mu_w = \frac{\|W(Ax - b)\|}{\|Ax - b\|} \quad (2.48)$$

Giving the expression for x

$$x = \begin{cases} A^\dagger b & \text{if } Ax = b \\ (WA)^\dagger Wb & \text{if } WAx = Wb \\ (A^T (W^T W + \mu_w I) A)^{-1} A^T (W^T W + \mu_w I) b & \text{otherwise} \end{cases} \quad (2.49)$$

where μ_w is defined in (2.48). The solution for the last one can be obtained through similar way as for the original least squares problem. In another variation of the problem the uncertainty can be given with respect to matrix A but in multiplicative form. Thus the robust optimization problem for this variation can be stated as follows:

$$\min_x \max_{\|\Delta A\| \leq \rho_A} \|(I + \Delta A)Ax - b\| \quad (2.50)$$

which can be reduced to the following minimization problem:

$$\min_x (\|Ax - b\| + \rho_A \|Ax\|) \quad (2.51)$$

then by similar analysis we obtain:

$$\frac{\partial \mathcal{L}_{\text{MLLS}}(x)}{\partial x} = \frac{A^T (Ax - b)}{\|A^T (Ax - b)\|} + \rho_A \frac{A^T Ax}{\|Ax\|} = 0 \quad (2.52)$$

and finally

$$x = \begin{cases} (A^T A)^\dagger b & \text{if } A^T Ax = A^T b \\ (A^T A (1 + \mu_A))^{-1} A^T b, \mu_A = \frac{\|A^T (Ax - b)\|}{\|Ax\|} & \text{otherwise} \end{cases} \quad (2.53)$$

2.6.1 Uncoupled Uncertainty

In the case that we have specific knowledge for the uncertainty bound of each data point separately we can consider the corresponding problem. The solution for this type of uncertainty reveals a very interesting connection between robustness and LASSO regression. Originally this result was obtained by Xu et al. [63]. Let us consider the least squares problem where the uncoupled uncertainties exist only with respect to the rows of the data matrix A :

$$\min_x \max_{\Delta A \in \mathcal{A}} \|(A + \Delta A)x - b\|_2 \quad (2.54)$$

where the uncertainty set \mathcal{A} is defined by:

$$\mathcal{A} \triangleq \{(\delta_1, \delta_2, \dots, \delta_m) \mid \|\delta_i\| \leq \rho_i\} \quad (2.55)$$

For the maximization problem and for a fixed vector x :

$$\max_{\Delta A \in \mathcal{A}} \|(A + \Delta A)x - b\|_2 = \max_{\Delta A \in \mathcal{A}} \|Ax - b + \Delta x\|_2 \quad (2.56)$$

$$= \max_{\Delta A \in \mathcal{A}} \|Ax - b + \sum_{i=1}^m x_i \delta_i\|_2 \quad (2.57)$$

$$\leq \max_{\Delta A \in \mathcal{A}} \|Ax - b\|_2 + \sum_{i=1}^m \|x_i \delta_i\|_2 \quad (2.58)$$

$$\leq \max_{\Delta A \in \mathcal{A}} \|Ax - b\|_2 + \sum_{i=1}^m |x_i| \cdot \rho_i \quad (2.59)$$

This provides an upper bound for the objective function. This bound is obtained by proper use of the triangular inequality. On the other side if we let

$$u = \begin{cases} \frac{Ax-b}{\|Ax-b\|_2} & \text{if } Ax \neq b \\ \text{any unit norm vector} & \text{otherwise} \end{cases} \quad (2.60)$$

Next we define the perturbation being equal to

$$\delta_i^* \triangleq \begin{cases} -c_i \cdot \text{sign}(x)u & \text{if } x_i \neq 0 \\ -c_i u & \text{o/w} \end{cases} \quad (2.61)$$

This perturbation belongs to the set of admissible perturbations since $\|\delta_i^*\|_2 = c_i$. If we set the perturbation in the maximization problem of (2.54) equal to (2.61), we get:

$$\max_{\Delta A \in \mathcal{A}} \|(A + \Delta A)x - b\|_2 \geq \|(A + \Delta A)x - b\|_2 \quad (2.62)$$

$$= \|(A + (\delta_1^*, \delta_2^*, \dots, \delta_m^*))x - b\|_2 \quad (2.63)$$

$$= \|Ax - b + \sum_{i: x_i \neq 0} (-x_i \cdot \text{sgn}(x_i)u)\|_2 \quad (2.64)$$

$$= \|Ax - b + u \cdot \sum_{i=1}^m c_i |x_i|\|_2 \quad (2.65)$$

$$= \|Ax - b\|_2 + \sum_{i=1}^m c_i |x_i| \quad (2.66)$$

The last equation combined with (2.59) yields that the maximization problem:

$$\max_{\Delta A \in \mathcal{A}} \|(A + \Delta A)x - b\|_2 \quad (2.67)$$

attains its maximum for the point $\Delta A = (\delta_1, \delta_2, \dots, \delta_m)$ where $\delta_i, i = 1, \dots, m$ is defined by (2.61). This proves that the original problem can be written as:

$$\min_x \max_{\Delta A \in \mathcal{A}} \|(A + \Delta A)x - b\|_2 = \min_x \left\{ \|Ax - b\|_2 + \sum_{i=1}^m c_i \cdot |x_i| \right\} \quad (2.68)$$

The last is nothing but a regularized linear least squares problem with l_1 regularization term. The last relation not only proves another interesting connection between regularization and robustness but also suggests a practical method for adjusting the regularization parameter in case that we have prior knowledge of data uncertainty.

As pointed out by the authors in [63] the above result can be generalized for any arbitrary norm. Thus the robust regression problem

$$\min_x \max_{\Delta A \in \mathcal{U}_p} \|(A + \Delta A)x - b\|_p \quad (2.69)$$

with

$$\mathcal{U}_p \triangleq \{(\delta_1, \delta_2, \dots, \delta_m) \mid \|\delta_i\|_p \leq \rho_i\} \quad (2.70)$$

is equivalent to the following problem

$$\min_x \left\{ \|Ax - b\|_p + \sum_{i=1}^m c_i \cdot |x_i| \right\} \quad (2.71)$$

This shows that LASSO type regularization can be the robust equivalent of a general regression problem regardless the norm given that the induced perturbations are defined as in (2.70).



<http://www.springer.com/978-1-4419-9877-4>

Robust Data Mining

Xanthopoulos, P.; Pardalos, P.; Trafalis, T.B.

2013, XII, 59 p. 6 illus., Softcover

ISBN: 978-1-4419-9877-4