AN INFO-GAP APPROACH TO LINEAR REGRESSION

M. Zachsenhouse, S. Nemets, A. Yoffe, Y. Ben-Haim

Faculty of Mechanical Engineering
Technion—Israel Institute of Technology
Haifa 32000 Israel

Mikhail A. Lebedev, Miguel A.L. Nicolelis

Dept. of Neurobiology and
Center for Neuro-engineering
Duke University, Durham NC, USA

ABSTRACT

Linear regression with high uncertainties in the measurements, model structure and model permanence is a major challenging problem. Standard regression techniques are based on optimizing a certain performance criterion, usually the mean squared error, and are highly sensitive to uncertainties. Regularization methods have been developed to address the problem of measurement uncertainty, but choosing the regularization parameter under severe uncertainties is problematic. Here we develop an alternative regression methodology based on satisficing rather than optimizing the performance criterion while maximizing the robustness to uncertainties. Uncertainties are represented by info-gap models which entail an unbounded family of nested sets of measurements parameterized by a non-probabilistic horizon of uncertainty. We prove and demonstrate that the robust-satisficing solution is different from the optimal least squares solution and that the info-gap approach can provide higher robustness to uncertainty.

1. INTRODUCTION

Linear regression analysis is based on a set of input/output observations which are expected to be linearly related. While linear regression methods are highly developed, their applicability to problems with severe uncertainties is still a challenge. Uncertainties may arise from three major sources: (i) measurement uncertainty, which accounts for inaccuracies in the observations, (ii) model uncertainty, which accounts for possible non-linear effects, and (iii) structural uncertainty, which accounts for potential changes over time in the parameters of the linear model or appearance of non-linear components. Standard linear regression techniques are based on optimizing a performance criterion, usually the mean squared error, but may fail to provide high, or even acceptable, performance for new observations. When the condition number of the measurement matrix is high, the least squares solution is greatly affected by uncertainties in the measurements and may differ considerably from the underlying linear model. Regularization methods have been developed to systematically reduce the condition number, but the choice of the regularization parameter is not easy under severe uncertainties.

Linear regression is usually performed to derive a model that can be used to make predictions of the future, and should therefore be designed to confront future surprises. Surprises are data generating processes differing from the processes which produced the data used to estimate the model. Surprises are structural changes in the system which arise from market changes, from technological innovations, or from discoveries of any sort. By definition, nothing in historical data can reveal anything about future surprises.

Surprises are a "true uncertainty" in the sense of Frank Knight [5] who was the first to make a clear distinction between risk, which involves known probability measures, and what Knight called true uncertainty where the underlying distributions are unknown.

In this paper we use info-gap models to represent uncertainty [1]. An info-gap model is an unbounded family of nested sets of events. There are no measure functions and there is no worst case. Info-gap models provide a stark and minimalistic representation of ignorance of future surprises.

A basic theorem of info-gap theory establishes a trade-off between fidelity to the historical data and robustness against info-gaps. Furthermore, maximization of fidelity (e.g. least squares estimation) has zero robustness to info-gaps. Robustness is obtained only by relinguishing fidelity. These properties are demonstrated in three theorems and an example.

2. LINEAR REGRESSION OF HIGHLY UNCERTAIN DATA

The regression. Consider the linear regression:

$$y_i = c^T x_i, \quad i = 1, \ldots, K, \quad y_i \in \Re^1, \; x_i \in \Re^N \qquad (1)$$

Denote $y = (y_1, \ldots, y_K)^T$ and $X = [x_1, \ldots, x_K]$. Eqs.(1) can now be succintly represented as $X^T c = y$. The mean squared error with regression coefficients c is:

$$S(y, X, c) = \frac{1}{K} \sum_{i=1}^{K} (y_i - c^T x_i)^2 \qquad (2)$$

Our observations are $\widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_K)^T$, $\widetilde{X} = [\widetilde{x}_i, \ldots, \widetilde{x}_K]$. We would like to choose c so that $S(\widetilde{y}, \widetilde{X}, c)$ is small and robust to info-gaps in the data.

Consider the following info-gap model for data variability:

$$\mathcal{U}(\alpha, \widetilde{y}, \widetilde{X}) = \Big\{ y, X : \ |y_i - \widetilde{y}_i| \le \alpha v_i, \qquad (3)$$
$$(x_i - \widetilde{x}_i)^T W_i^{-1} (x_i - \widetilde{x}_i) \le \alpha^2, \ i = 1, \ \ldots, \ K \Big\}$$

where $\alpha \ge 0$, $v_i > 0$ and W_i is real, symmetric and positive definite. v_i and W_i are known.

Formulating the robustness. The robustness of regression c, with critical squared error S_c, is the greatest horizon of uncertainty α up to which all data realizations have squared error no greater than S_c:

$$\widehat{\alpha}(c, S_c) = \sup \left\{ \alpha : \ \left(\sup_{y, X \in \mathcal{U}(\alpha, \widetilde{y}, \widetilde{X})} S(y, X, c) \right) \le S_c \right\} \qquad (4)$$

Denote the inner maximum in this expression by $M(\alpha)$. $M(\alpha)$ increases as α increases, and the robustness is the greatest value of α at which $M(\alpha) \le S_c$. If $M(\alpha)$ is strictly monotonic then it is the inverse of $\widehat{\alpha}(c, S_c)$:

$$M(\alpha) = S_c \quad \text{if and only if} \quad \widehat{\alpha}(c, S_c) = \alpha \qquad (5)$$

The robust-satisficing regression, $\widehat{c}(S_c)$, maximizes the robustness and satisfices the performance:

$$\widehat{c}(S_c) = \arg\max_c \widehat{\alpha}(c, S_c) \qquad (6)$$

The nominally optimal regression, $c^\star$, minimizes the mean-squared error based on the data:

$$c^\star = \arg\min_c S(\widetilde{y}, \widetilde{X}, c) \qquad (7)$$

Evaluating the robustness. $M(\alpha)$ is obtained when each of the terms $(y_i - c^T x_i)^2$ in the sum in eq.(2) is maximal, at horizon of uncertainty α. The maximum of $(y_i - c^T x_i)^2$ occurs when y_i and $c^T x_i$ are extremal: one max and the other min. The extremal values of y_i and $c^T x_i$ are:

$$\max_{y, X \in \mathcal{U}(\alpha, \widetilde{y}, \widetilde{X})} y_i \ = \ \widetilde{y}_i \pm \alpha v_i \qquad (8)$$

$$\max_{y, X \in \mathcal{U}(\alpha, \widetilde{y}, \widetilde{X})} c^T x_i \ = \ c^T \widetilde{x}_i \pm \alpha \sqrt{c^T W_i c} \qquad (9)$$

From this one finds:

$$M(\alpha) = \frac{1}{K} \sum_{i=1}^{K} \left[|\widetilde{y}_i - c^T \widetilde{x}_i| + \left(v_i + \sqrt{c^T W_i c} \right) \alpha \right]^2 \qquad (10)$$

$$= F\alpha^2 + G\alpha + \widetilde{S} \qquad (11)$$

$\widetilde{S} = S(\widetilde{y}, \widetilde{X}, c)$ is the mean squared error of the data, and F and G are both positive and defined as:

$$F \ = \ \frac{1}{K} \sum_{i=1}^{K} \left(v_i + \sqrt{c^T W_i c} \right)^2 \qquad (12)$$

$$G \ = \ \frac{2}{K} \sum_{i=1}^{K} |\widetilde{y}_i - c^T \widetilde{x}_i| \left(v_i + \sqrt{c^T W_i c} \right) \qquad (13)$$

Consider the quadratic polynomial:

$$F\alpha^2 + G\alpha + \widetilde{S} - S_c = 0 \qquad (14)$$

The robustness is zero if $\widetilde{S} - S_c > 0$. Otherwise the robustness is the lowest non-negative root. From the relation between roots, α_1 and α_2, and coefficients we know that $(\widetilde{S} - S_c)/F = \alpha_1 \alpha_2$. Thus the roots are of opposite sign and the robustness is:

$$\widehat{\alpha}(c, S_c) = \frac{1}{2F} \left(-G + \sqrt{G^2 + 4F(S_c - \widetilde{S})} \right) \qquad (15)$$

3. CROSSING OF ROBUSTNESS CURVES

Proposition 1 Given: *$\widetilde{y}$ and $\widetilde{X}$ are not identically zero.*

Then: *There exist a vector c different from $c^\star$ and a value of S_c such that $\widehat{\alpha}(c, S_c) = \widehat{\alpha}(c^\star, S_c)$.*

Proof of proposition 1: According to (5) it is enough to show that there exist a vector c and a value of α such that $M(c, \alpha) = M(c^\star, \alpha)$.

We consider the function $P(c, \alpha) = M(c, \alpha) - M(c^\star, \alpha)$. In light of eq.(11), $P(c, \alpha)$ is a parabola in α whose coefficients are functions of c. That is:

$$P(c, \alpha) = A(c)\alpha^2 + B(c)\alpha + C(c) \qquad (16)$$

were:

$$A(c) \ = \ F(c^\star) - F(c), \ B(c) = G(c^\star) - G(c) \qquad (17)$$

$$C(c) \ = \ \widetilde{S}(c^\star) - \widetilde{S}(c) \qquad (18)$$

and F, G, $\widetilde{S}$ are as defined in (11)–(13).

We note that $c^\star$ minimizes $\widetilde{S}$ so $\widetilde{S}(c^\star) \le \widetilde{S}(c)$ and $C(c) \le 0$ for any c. Thus the equation defined by $P(c, \alpha) = 0$ has a non-negative root $\alpha_\times$ when $A(c) > 0$. Consider $c = \delta c^\star$ where $0 < \delta < 1$. Then for each W_i we have:

$$c^T W_i c = \delta^2 c^{\star T} W_i c^\star < c^{\star T} W_i c^\star \qquad (19)$$

Consequently, from the definition of $F(c)$ in eq.(12) we find $F(\delta c^\star) < F(c^\star)$. Hence $A(\delta c^\star) > 0$. Thus the robustness curve for $c = \delta c^\star$ crosses the robustness curve for $c^\star$ for some non-negative root $\alpha_\times$. ∎

4. ROBUSTNESS OF REGULARIZED SOLUTIONS

Proposition 1 indicates that there are regression vectors which become more robust than the optimal regression $c^\star$ at some level of the satisfying mean squared error (MSE) S_c. In this section we evaluate the robust-satisfying properties of specific sequences of parameterized regression vectors, and show that they become more robust than the optimal regression $c^\star$ at some level of satisficing MSE S_c.

4.1. Optimal Least Square Solution

The SVD of the observed matrix $\widetilde{X}^T$ is given by:

$$\widetilde{X}^T = \widetilde{U}^T \widetilde{\Sigma} \widetilde{V} = \sum_{i=1}^{r} \widetilde{u}_i \sigma_i \widetilde{v}_i \tag{20}$$

where $\widetilde{U} = [\widetilde{u}_i, \ldots, \widetilde{u}_K] \in \Re^{K \times K}$ and $\widetilde{V} = [\widetilde{v}_i, \ldots, \widetilde{v}_K] \in \Re^{M \times N}$ are ortho-normal matrices, $\widetilde{\Sigma} = \text{diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_N) \in \Re^{N \times N}$ with $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \cdots \geq \widetilde{\sigma}_N \geq 0$, and the rank $r \leq N$ is the number of strictly positive singular values $\widetilde{\sigma}_i$.

Since $\widetilde{U}^T \widetilde{U} = I_{K \times K}$ the correlation matrix of the data is given by:

$$\widetilde{R} = \widetilde{X} \widetilde{X}^T = \widetilde{V} \widetilde{\Sigma}^2 \widetilde{V}^T \in \Re^{N \times N} \tag{21}$$

The least-squares (LS) solution can be expressed in terms of the SVD as:

$$c_{\text{LS}} = \sum_{i=1}^{r} \frac{\widetilde{u}_i^T \widetilde{y}}{\widetilde{\sigma}_i} \widetilde{v}_i \tag{22}$$

The resulting MSE is given by:

$$\widetilde{S}_{\text{LS}} = \widetilde{S}(c_{\text{LS}}) \left\| \widetilde{y} - \widetilde{X}^T c_{\text{LS}} \right\|^2 = \sum_{i=1}^{K} (\widetilde{u}_i^T \widetilde{y})^2 \tag{23}$$

The LS optimal solution is highly sensitive to measurement errors since, from eq.(22), it depends on the inverse of the singular values. Small singular values can dominate the solution and magnify errors in the measurement vector y. Thus it is common to use regularization to stabilize the solution.

4.2. Crossing of Robustness Curves for Regularized Least Square Regression

We consider two classes of parameterized regression vectors obtained by: (a) Truncated SVD, and (b) Tikhonov regularization [3].

4.2.1. Truncated Least Squares

The truncated-LS regression is obtained by truncating the LS regression of eq.(22) at $k \leq r \leq N$:

$$c_k = \sum_{i=1}^{k} \frac{\widetilde{u}_i^T \widetilde{y}}{\widetilde{\sigma}_i} \widetilde{v}_i \tag{24}$$

The resulting MSE is:

$$\widetilde{S}_k = \widetilde{S}(c_k) = \left\| \widetilde{y} - \widetilde{X}^T c_k \right\|^2 = \sum_{i=k+1}^{K} (\widetilde{u}_i^T \widetilde{y})^2 \tag{25}$$

$$= \widetilde{S}_{\text{LS}} + \sum_{i=k+1}^{r} (\widetilde{u}_i^T \widetilde{y})^2 \tag{26}$$

which increases as more terms in the LS regression are truncated.

Proposition 2 Given: *the info-gap model of eq.(3) with $W_i = W$ and $v_i = \mu$, for all i.*

Then: *The robustness curves of all the different truncated-LS regression vectors cross each other when the eigenvectors of the shape matrix W are the same as the eigenvectors of the correlation matrix of the observations $\widetilde{R}$, i.e., $W = \widetilde{V} \Sigma_W^2 \widetilde{V}^T$ where $\Sigma_W = diag(\sigma_{W_1}, \ldots, \sigma_{W_N})$ where σ_{W_i} is the ith singular value of W.*

The proof of proposition 2 is not presented as it is similar to the proof of proposition 3 which is included.

Note that the following weight matrices are special cases of the weight matrix considered in proposition 2: (a) the correlation matrix of the observations $W = \widetilde{R}$, (b) the inverse of the correlation matrix $W = \widetilde{R}^{-1}$, (c) the identity matrix $W = I$, or (d) a non-negative combination of the above, i.e., $W = \beta_0 I + \beta_1 \widetilde{R} + \beta_2 \widetilde{R}^{-1}$ with $\beta_1, \beta_2, \beta_3 \geq 0$.

4.2.2. Tikhonov Regularization

Tikhonov regularization stabilizes the optimal LS solution by minimizing the combination of the MSE and the size of the regression vector: $\min \left[\left\| \widetilde{y} - \widetilde{X}^T c \right\|^2 + \lambda^2 \|Lc\|^2 \right]$. When $L = I$ the regularized regression vector with a given λ^2 is:

$$c_\lambda = \sum_{i=1}^{r} \frac{\widetilde{\sigma}_i^2}{\widetilde{\sigma}_i^2 + \lambda^2} \frac{u_i^T y}{\sigma_i} v_i \tag{27}$$

The resulting MSE is:

$$\widetilde{S}_\lambda = \widetilde{S}(c_\lambda) = \left\| \widetilde{y} - \widetilde{X}^T c_\lambda \right\|^2 \tag{28}$$

$$= \sum_{i=1}^{r} \frac{\lambda^4}{(\widetilde{\sigma}_i^2 + \lambda^2)^2} (\widetilde{u}_i^T \widetilde{y})^2 + \sum_{i=r+1}^{K} (\widetilde{u}_i^T \widetilde{y})^2 \tag{29}$$

Note that:

$$\frac{\partial \widetilde{S}_\lambda}{\partial \lambda^2} = 2\lambda^2 \sum_{i=1}^{r} \frac{\widetilde{\sigma}_i^2}{(\widetilde{\sigma}_i^2 + \lambda^2)^3} (u_i^T y)^2 > 0 \tag{30}$$

Hence the MSE increases with λ^2 and the minimum is achieved for the optimal LS solution, $c_{\lambda=0} = c_{\text{LS}}$.

Proposition 3 Given: *the info-gap model of eq.(3) with $W_i = W$ and $v_i = \mu$, for all i.*

Then: *The robustness curves of any two regularized regression vectors $c_{\lambda_1} \neq c_{\lambda_2}$ cross each other when the eigenvectors of the shape matrix W are the same as the eigenvectors of the correlation matrix of the observations $\widetilde{R}$, i.e., $W = \widetilde{V} \Sigma_W^2 \widetilde{V}^T$ where $\Sigma_W = diag(\sigma_{W_1}, \ldots, \sigma_{W_N})$ where σ_{W_i} is the ith singular value of W.*

Proof of proposition 3. According to eq.(5) it is enough to show that there is a non-negative α such that $M(c_{\lambda_1}, \alpha) = M(c_{\lambda_2}, \alpha)$. $c_{\lambda_1} \neq c_{\lambda_2}$ implies $\lambda_1 \neq \lambda_2$ so with no loss of generality we assume that $\lambda_1 < \lambda_2$. Hence, according to eq.(30), $\widetilde{S}(c_{\lambda_1}) < \widetilde{S}(c_{\lambda_2})$, and thus $M(c_{\lambda_1}, \alpha = 0) < M(c_{\lambda_2}, \alpha = 0)$. Following the argument in the proof of proposition 1, this relationship is reversed for some positive α if $F(c_{\lambda_1}) > F(c_{\lambda_2})$ where, from eq.(12), $F(c_\lambda) = \left(\mu + \sqrt{c_\lambda^T W c_\lambda}\right)^2$.

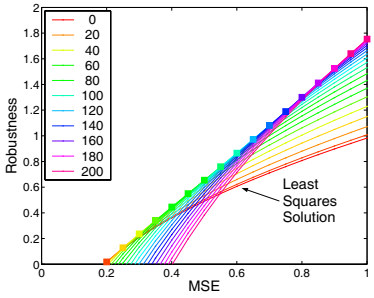
Hence we need to evaluate $c_\lambda^T W c_\lambda$ which can be expressed:

$$c_\lambda^T W c_\lambda = c_\lambda^T \widetilde{V} \Sigma_W^2 \widetilde{V}^T c_\lambda = \sum_{i=1}^{r} \frac{\widetilde{\sigma}_i^2}{\widetilde{\sigma}_i^2 + \lambda^2} \left(\frac{\widetilde{u}_i^T \widetilde{y}}{\widetilde{\sigma}_i} \sigma_{W_i} \right)^2 \tag{31}$$

The derivative of each term in eq.(31) with respect to λ^2 is negative. Thus $c_\lambda^T W c_\lambda$ decreases as λ^2 increases and $c_{\lambda_1}^T W c_{\lambda_1} > c_{\lambda_2}^T W c_{\lambda_2}$. Consequently, $F(c_{\lambda_1}) = \left(\mu + \sqrt{c_{\lambda_1}^T W c_{\lambda_1}}\right)^2 > \left(\mu + \sqrt{c_{\lambda_2}^T W c_{\lambda_2}}\right)^2 = F(c_{\lambda_2})$. From this we conclude that $M(c_{\lambda_1}, \alpha) = M(c_{\lambda_2}, \alpha)$ has a positive for positive α. ∎

5. EXAMPLE

We demonstrate the robustness-performance trade-off for a linear regression problem based on neuronal data. Neural spike trains can be modeled as doubly stochastic Poisson processes [4, 7], where the underlying rate is a stochastic process that depends on behaviorally relevant signals. A major interest is the ability to "read the neural code", i.e., to reconstruct the behavioral signals that are encoded in the neural activity from neural activity of an ensemble of neurons. The specific application considered here is the reconstruction of the velocity of movement from the neural activity recorded from a large ensemble of neurons (183 in the present example) and is based on the experiments reported in [2, 6]. Fig. 1 depicts the robustness curves of different regularized solution for $\lambda = 0$ to $\lambda = 200$. All the robustness curves cross each other and in particular they cross the robustness curve of the least squares solution. Thus, increasing robustness to uncertainties can be achieved by selecting regularized solutions with increasing levels of the regularization parameter, at the expense of decreasing fidelity to the data. At any satisficing performance level, specified by a MSE higher than the optimum, the regularization parameter providing the highest robustness can be selected.



6. REFERENCES

1. Ben-Haim, Y., *Information-gap Decision Theory: Decisions Under Severe Uncertainty,* Academic Press, San Diego (2001).
2. Carmena, J.M., M.A. Lebedev, R.E. Crist, J.E. O'Doherty, D.M. Santucci, D. Dimitrov, P.G. Patil, C.S. Henriquez, M.A.L. Nicolelis, Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol.*, 1, 193-208, (2003).
3. Golub, G.H. & Van-Loan, C.F., *Matrix Computations,* The Johns Hopkins University Press, (1996).
4. Johnson, D.H., Point Process models of single neuron discharge. *J. Comp. Neurosci.*, 3, 275-299, (1996).
5. Knight, Frank H., 1921, *Risk, Uncertainty and Profit,* Houghton Mifflin Co. Re-issued by University of Chicago Press, 1971.
6. Lebedev, M.A., J.M. Carmena, J.E. O'Doherty, M. Zacksenhouse, C.S. Henriquez, J.C. Principe, M.A.L. Nicolelis, Cortical ensemble adaptation to represent velocity of an artificial actuator controlled by a brain machine interface, *J. Neurosci.*, 25(19), 4681-4693 (2005).
7. Snyder, D.L., *Point Processes,* John Wiley & Sons, Inc., (1975).