Regression to a Linear Lower Bound With Outliers: An Exponentially Modified Gaussian Noise Model

Julien Gori, Olivier Rioul

Regression to a Linear Lower Bound With Outliers: An Exponentially Modified Gaussian Noise Model

Julien Gori
LRI, Univ. Paris-Sud, CNRS
Inria, Université Paris-Saclay
F-91400, Orsay, France

Olivier Rioul
LTCI, Télécom Paris
Institut Polytechnique de Paris
F-75013, Paris, France

***Abstract*—A regression method to estimate a linear bound in the presence of outliers is discussed. An exponentially-modified Gaussian (EMG) noise model is proposed, based on a maximum entropy argument. The resulting "EMG regression" method is shown to encompass the classical linear regression (with Gaussian noise) and a minimum regression (with exponential noise) as special cases. Simulations are performed to assess the consistency of the regression as well as its resilience to model mismatch. We conclude with an example taken from a real-world study of human performance in rapid aiming with application to human-computer interaction.**

I. Introduction

The aim of this work is to present a general method to estimate linear lower bounds in two-dimensional datasets (scatter plots) in the presence of outliers. Consider a set of n observed samples (x_i, y_i), $i = 1, 2, \ldots, n$, which are independent and identically distributed (i.i.d.) realizations of an input variable X and output variable Y. A typical example is represented in Fig. 1. We assume the following characteristics:

- X is the independent variable and is perfectly known;
- Y is the dependent variable, subject to observation errors;
- $\min Y|(X = x) = a + b\ x$ holds when there are no observation errors at all. In words, a linear lower bound exists in the dataset.
- Some observations (outliers) go actually below the lower bound.

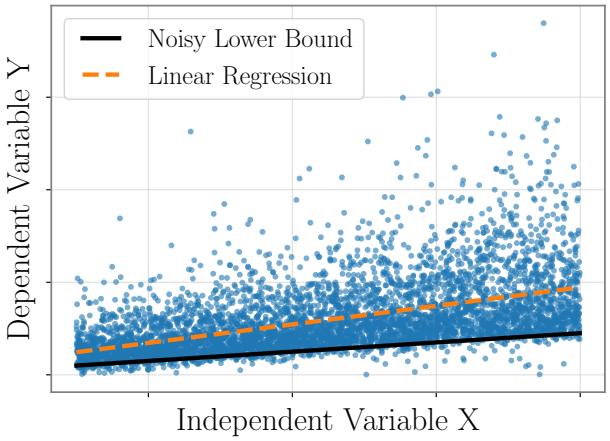

Fig. 1. Example dataset with a lower bound (black straight line) to be estimated. Standard linear regression gives the orange dashed line.

As a result, the observed conditional probability densities $p_{Y|X=x}$ are positively skewed. The proposed method allows regression towards a lower bound, while taking into account both heavily skewed distributions and the presence of outliers.

Existing Regression Techniques: There exist many regression techniques, the best known being the standard linear regression. The linear regression method assumes that (i) X is perfectly known; (ii) the relationship between X and Y is linear in the sense that $\mathbb{E}[Y|X = x] = a + b\ x$ where a is the intercept and b is the slope of the linear model; (iii) observation errors are accounted for by the probabilistic model

$$Y = a + b\ X + Z, \tag{1}$$

where Z denotes the model noise. It is usually assumed that errors are centered, uncorrelated, with the same finite variance (homoscedastic). In this case, the Gauss-Markov theorem states that the ordinary least squares (OLS) estimator is the best linear unbiased estimator. This estimator is easily tractable analytically, and the parameters to be estimated are simple to compute, making simple linear regression[1] a very popular tool among both novices and experts.

Despite all these attractive features, linear regression is not a silver bullet. Estimated parameters are oversensitive to outliers [1] and OLS may be outperformed by the minimum absolute deviations estimator [2]. Techniques exist to account for non-Gaussian distributions, by transforming the data prior to regression [3], by using mixture models [4], or generalized linear models [5]. Common to all these methods, however, is that the regression is designed to assess a *central* tendency. In contrast, the aim of this paper is to estimate a *lower* bound.

Extreme value theory (EVT) provides a counterpart to the central limit theorem for extrema rather than means through the Fisher-Tippett-Gnedenko theorem [6]. Such a theoretical result seems at first sight adapted to the present work. However, EVT essentially describes the behavior of tail events, and using it in our case would amount to fitting the outliers below the lower bound. Instead, our present work focuses on estimating the linear lower bound $a + bx$.

[1]We refer here to linear regression as a method by which a and b are computed using OLS. If one further assumes that Z is Gaussian, it is well known that the OLS estimator is equivalent to the maximum likelihood estimator.

To the best of our knowledge, previous regression techniques do not solve the problem of estimating a linear lower bound in the presence of outliers (the black straight line of Fig. 1).

Outline of the Paper: The remainder of this paper is as follows. Section II describes an exponentially-modified Gaussian (EMG) noise model and shows how it is adapted to our problem formulation. Section III then describes the parametric maximum likelihood (ML) estimation leading to the desired linear lower bound. Two limit cases are discussed. Simulations to assess the validity and robustness of the methods are conducted in Section IV. A real-world example is presented in Section V, upon which EMG regression is applied. Section VI concludes.

II. Exponentially-Modified Gaussian Model

The proposed model for the conditional $Y|X$ is a linear model with two independent additive noise components:

$$Y = a + b\ X + E + Z, \tag{2}$$

where Z is a centered Gaussian noise with variance σ^2 and density

$$p_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \tag{3}$$

and where $E \geq 0$ is a nonnegative random variable. Each term of (2) is interpreted as follows:

- a and b are the two parameters that specify the linear lower bound,
- $E \geq 0$ accounts for the fact that observations can in principle only be above the lower bound—it captures most of the observations.
- Z accounts for symmetric deviations from the lower bound model—in particular, it captures *outliers* below the lower bound.

In datasets such as the one illustrated in Fig. 1, it is expected that the E scales much larger than the deviation σ of Z. This is in essence due to the fact that the outliers on the bottom (captured by Z) are much closer to the lower bound than those at the top (captured by E) where Y is completely unconstrained.

We now discuss how the distribution of E can be chosen so as to retain a sense of generality to the regression method, rather than to reflect an underlying "true" model. We rely on the maximum entropy (MaxEnt) principle [7] which assigns probability density functions "in a way that is maximally noncommital" given the observed data and constraints thereupon. Jaynes [7] has shown that this statement can be formulated as a constrained optimization problem where Shannon's entropy [8] is maximized given the constraints.

It is reasonable to assume that the mean value of E given $X = x$, say β, exists and is given. It is easily seen [9, Ex. 12.2.5] that the MaxEnt distribution of a *positive* random variable with fixed mean β is the exponential distribution of parameter β:

$$p_E(t) = \frac{1}{\beta} \exp\left(-\frac{t}{\beta}\right). \tag{4}$$

The sum of independent Gaussian and exponential random variables is known to follow the exponentially-modified Gaussian distribution [10]:

$$p(y|x) = \frac{1}{2\beta} e^{\frac{1}{2\beta}(2\mu + \frac{\sigma^2}{\beta} - 2y)} \operatorname{erfc}\left(\frac{\mu + \frac{\sigma^2}{\beta} - y}{\sqrt{2}\sigma}\right) \tag{5}$$

where $\mu = a + b\ x$ and erfc is the complementary error function. EMG distributions have been used to describe highly skewed shapes of peaks in chromatography [11] and of distributions for recall/reaction time in experimental psychology [10]. The EMG distribution is illustrated Fig. 2 for different values of σ^2 with $\mu = 0$ and $\beta = 1$. Not surprisingly, the EMG distribution converges to the exponential one when $\sigma^2 \to 0$, and to the Gaussian one when $\beta \to 0$ (see Section III).

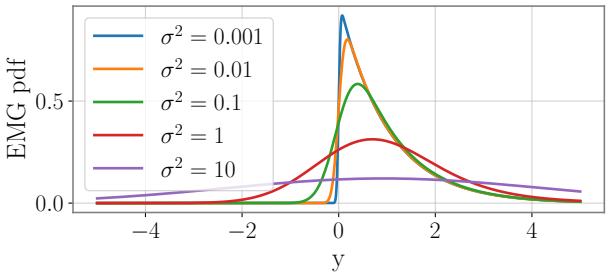

Fig. 2. EMG distribution (5), with $\mu = 0$, $\beta = 1$, and various values of σ^2

III. The EMG Regression Method

In our proposed regression method, the parameters a, b of (2) are determined through maximum likelihood estimation—which under mild conditions is known to be asymptotically consistent and efficient. The log-likelihood function $\ell = \ell(a, b, \beta, \sigma^2)$ associated with model (5) for n i.i.d. samples (x_i, y_i) is

$$\ell = -n \log 2\beta + \frac{1}{2\beta} \sum_{i=1}^{n} \left(2(a + b\ x_i) + \frac{\sigma^2}{\beta} - 2y_i\right) + \sum_{i=1}^{n} \log \operatorname{erfc}\left(\frac{a + b\ x_i + \frac{\sigma^2}{\beta} - y_i}{\sqrt{2\sigma^2}}\right). \tag{6}$$

Its maximum is found numerically using well-known computational methods (see Section IV-A below).

Two Extreme Cases of the EMG Regression: By making the exponential component vanish (letting $\beta \to 0$ so that E tends to 0 in distribution), the EMG regression should normally result in a linear regression with Gaussian noise only, which amounts to a standard linear regression. Similarly, making the Gaussian component vanish (by letting $\sigma^2 \to 0$ so that Z tends to 0 in distribution), the EMG regression results in a regression with exponential noise only. These two extreme cases are calculated as follows.

a) $\beta \to 0$*:* Write $\lambda = \frac{1}{\beta} \to +\infty$. Using the asymptotic expansion $\operatorname{erfc}(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}(\frac{1}{x} + o(\frac{1}{x^2}))$ for $x \to +\infty$, one has $\sum_{i=1}^n \log \operatorname{erfc}\left(\frac{\mu+\lambda\sigma^2-y_i}{\sqrt{2}\sigma}\right) = \sum_{i=1}^n -\frac{(\mu-y_i)^2}{2\sigma^2} - \frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2y_i) + \sum_{i=1}^n \log(\frac{\sqrt{2}\sigma}{\sqrt{\pi}(\mu+\lambda\sigma^2-y_i)} + o(\frac{1}{\lambda^2}))$. The second term in the expansion cancels in (6). Taking the limit we obtain

$$\lim_{\beta\to 0} \operatorname*{argmax}_{a,b}(\ell) = \operatorname*{argmin}_{a,b} \sum_{i=1}^{n} (\mu - y_i)^2 \tag{7}$$

Therefore, as $\beta \to 0$, the EMG regression becomes equivalent to the classical (OLS) linear regression.

b) $\sigma^2 \to 0$*:* One has

$$\lim_{\sigma^2\to 0} p(y|x) = \frac{1}{2\beta} e^{\frac{1}{\beta}(\mu-y)} \lim_{\sigma^2\to 0} \operatorname{erfc}\left(\frac{\mu-y}{\sqrt{2}\sigma}\right), \tag{8}$$

where the limit of $\operatorname{erfc}\left(\frac{\mu-y}{\sqrt{2}\sigma}\right)$ equals 0 if $\mu - y > 0$ and equals 2 if $\mu - y < 0$. Therefore,

$$\lim_{\sigma^2\to 0} p(y|x) = \frac{1}{\beta} e^{-\frac{1}{\beta}(y-\mu)} \mathbb{1}_{y\geq\mu} \tag{9}$$

which is an exponential density with log-likelihood

$$\ell = -n \log \beta - \sum_{i=1}^{n} \frac{1}{\beta}(y_i - a - b\, x_i) \log \mathbb{1}_{y_i \geq a + b\, x_i} \tag{10}$$

Let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ be the average observed value of X. We may assume $\overline{x} > 0$, otherwise replace b by $-b$. Then (10) is increasing in both a and b under the condition $y \geq a + b\,x$. Hence maximizing likelihood amounts to maximizing $a + b\overline{x}$ for all a's and b's such that $y_i \geq a + b\, x_i$ for all i. In practice this can be achieved by finding all lines lying below the data that tangent the convex hull of the dataset, and keeping the one which maximizes $a + b\,\overline{x}$. Such a regression is determined by only two points, and is thus severely affected by outliers.

From the above two limiting cases we see that in essence, the EMG regression determines the lower bound from a combination of (symmetric) Gaussian and (asymmetric) exponential noises, where Gaussian noise implies a central tendency and where exponential noise implies a strict minimum.

IV. Simulations

Simulations were conducted to analyse the effect of the sample size and also to observe the effect of a model mismatch on the quality of the estimates. For each simulation, datasets of different sample sizes ($n \in [10, 21, 46, 100, 215, 464, 1000, 2154, 4641, 10000]$) were generated and fed, 500 times each, to the EMG regression.

A. Estimation Procedure

We used a global basin-hopping [12] optimizing method with 20 iterations, with a trust-exact method local optimizer [13] (both from the SciPy implementation) to find the extremum. The basin-hopping algorithm is a stochastic method that works by determining a new starting point at each iteration that the deterministic trust-exact method then uses. Although this gives no guarantee that the global extremum is achieved,

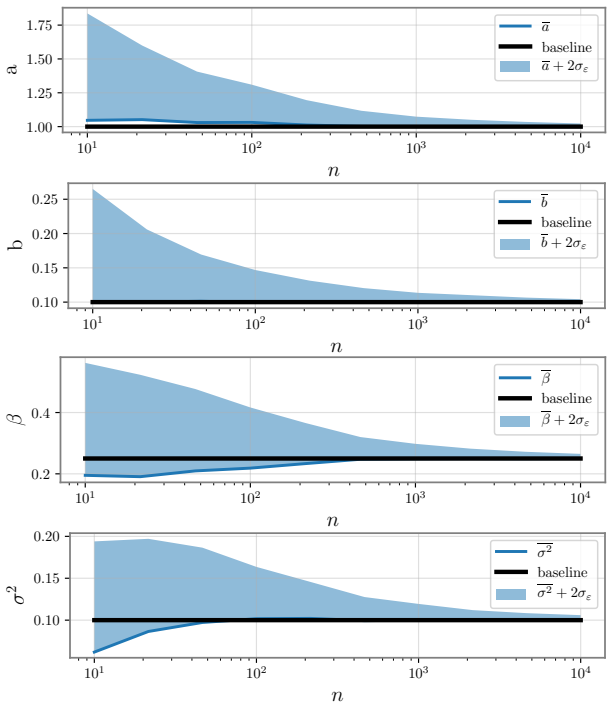

Fig. 3. Sample mean and sample standard deviation represented for each parameter $a = 1$, $b = 0.1$, $\beta = 1/4$, and $\sigma^2 = 0.1$. For each estimator t, $\overline{t}$ represents its sample average and $\sigma_{\overline{t}}$ is the associated sample standard deviation. The filled region represents $\hat{t} \in [\overline{t}, \overline{t} + 2\sigma_{\overline{t}}]$ (i.e., one side of the asymptotic 95% confidence area).

20 iterations seemed to be a good tradeoff between simulation time and effectiveness of the method to determine the global extremum. In order to decrease simulation time, the basin-hopping procedure was stopped before 20 iterations whenever a current minimum value was reached for the fifth time. Gradient and Hessian were straightforwardly computed for usage by the trust-exact method as numerical differentiation proved unreliable for small values of σ^2. We also used the scaled complementary Gaussian error function erfcx to avoid underflow. The starting point for the global algorithm was determined by a method of moments estimator[2] heuristic.

B. Consistency

We expect that the EMG regression method, based on ML estimation, is asymptotically consistent (the estimation bias vanishes) with a vanishing covariance matrix.

Our first simulation consisted of generating datasets according to (5) and observing how the estimates evolved with n. X was uniformly distributed and $Y|X$ was drawn according to (5), with parameters $a = 1$, $b = 0.1$, $\beta = 0.25$, $\sigma^2 = 0.1$[3].

[2]The method of moment estimator is known to be outperformed by the ML estimator, but its computation is very simple, see e.g. [14]. It is thus very useful to determine a starting point for the global optimization routine. The method of moments estimator will sometimes yield values outside of the admissible parameter space, such as negative values for positive parameters (e.g. σ^2). In that case, we arbitrarily replaced the estimated value by a very small one (say, 10^{-4}).

[3]The parameter values correspond approximately to those encountered in the real-world example of Section V.

We determined the mean and standard deviations σ_ε of the estimators for each parameter. Fig. 3 summarizes the results of the simulation. The estimate for b is unbiased for all values of n, while the estimates for a, β and σ^2 appear biased for small values of n but become unbiased in practice for a when $n \geq 200$, for β when $n \geq 500$ and for σ^2 when $n \geq 50$. The standard error decreases in n for all parameters; the coefficients of variation for a, b, β, σ^2 are respectively 1.1%, 2.1%, 3.0% and 3.0% for $n = 10000$. The estimators thus behave as expected.

C. Model Mismatch

Since the exponential distribution was chosen based on a generic MaxEnt argument, a natural question is whether or not the actual distribution for E in (2) affects the performance behavior of the estimation method. To answer this, we investigated a model mismatch, assuming a Weibull distribution for E rather than the exponential:

$$p_W(x) = \frac{s}{k}\left(\frac{x}{k}\right)^{s-1} \exp\left(-\left(\frac{x}{k}\right)^{s}\right) \qquad (x \geq 0), \qquad (11)$$

where k is the scale and s is the shape. For $s = 1$, the Weibull distribution (11) reduces to the exponential distribution (4) with $k = \beta$, while it is heavy-tailed for $s < 1$ and light-tailed for $s > 1$[4]. Also note that the mode of W shifts to the right for $s > 1$. Different kinds of mismatch can thus be chosen for different values of s.

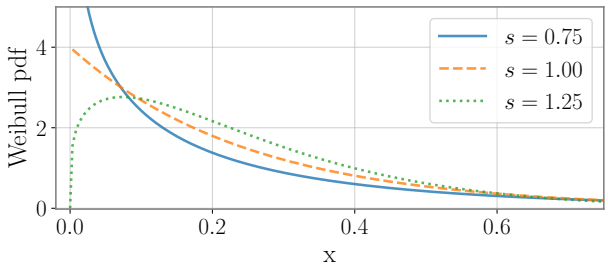

Fig. 4. Weibull probability density function for $k' = (4\Gamma(1+1/s))^{-1}$ and different values of s. For $s = 1$ this reduces to the exponential distribution with $\beta = k = 1/4$.

For a fair comparison, the parameters of the generating model were chosen so as to keep the first two moments of the scatter plot unchanged. Since the mean and variance of the Weibull distribution are [14] mean $\mu_0 = k\Gamma(1+1/s)$ and variance $V_0 = k^2\big(\Gamma(1+2/s) - \Gamma(1+1/s)^2\big)$, invariance of the first two moments is achieved by adjusting $k_0 = k/\Gamma(1+1/s)$ and replacing σ in (3) by $\sigma_0^2 = \sigma^2 + k^2\big(1 - \frac{\Gamma(1+2/s) - \Gamma(1+1/s)^2}{\Gamma(1+1/s)^2}\big)$. The Weibull distribution corrected with k' is illustrated Fig. 4 for values of shape $s \in [0.75, 1, 1.25]$.

Two simulations were conducted with $s = 0.75$ and $s = 1.25$. The results for $s = 0.75$ are given Fig. 5, those for $s = 1.25$ being similar. The baseline for β is $1/4$, as β should be equal to the mean; the baseline for σ^2 is the corrected σ_0^2. The

[4]A random variable X with a heavy-tailed distribution satisfies $\lim_{x\to\infty} \exp(tx)\Pr[X > x] = \infty$ for all $t > 0$.

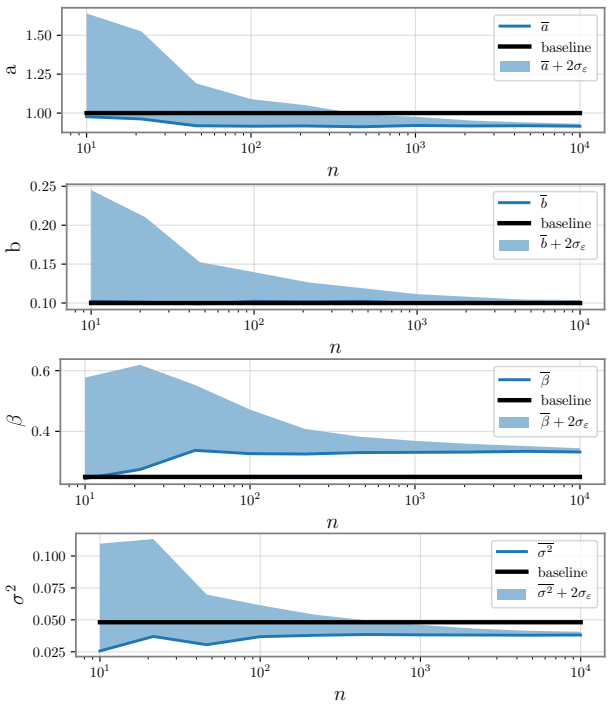

Fig. 5. Sample mean and sample standard deviation represented for each parameter when E is Weibull distributed. The parameters used to generate the data were $a = 1$, $b = 0.1$, $s = 0.75$, $k' = (\beta\Gamma(1 + 1/0.75))^{-1} = 0.210$, $\sigma^2 = 0.048$. For each estimator t, $\bar{t}$ represents its sample average and σ_ε is the associated sample standard deviation. The filled region represents $\hat{t} \in [\bar{t}, \bar{t} + 2\sigma_\varepsilon]$ (i.e., one side of the asymptotic 95% confidence area).

simulation shows that b is unbiased, while a, β, σ^2 are biased with respectively -0.1, $+0.08$, -0.01 of bias. The coefficients of variation for a, b, β, σ^2 are respectively 0.8%, 1.7%, 1.8%, 3.4% for $n = 10000$. The simulations for $s = 1.25$ indicated almost identical results except for sign changes of the biases. The sample variance being spread over β and σ^2, it is expected that an overestimation of one of these two parameters leads to an underestimation of the other, which in turn affects the estimated intercept a. Remarkably, the slope b is very robust to the model mismatch; the quality of the estimation of b was almost equivalent across all simulations.

V. A Real World Example

The motivation for this work stems from a real-world example in a study of human aiming performance. Existing theoretical work on human movement [15], [16] predicts that the shortest amount of time, say Y (in seconds), that is needed to successfully aim towards a target in a so-called Fitts task [17] is linearly related to a difficulty parameter X (in bits). Aiming data is usually gathered in a controlled experimental setting where participants are asked to minimize task completion time, and the parameters of the model are estimated through linear regression. Typical values for a and b in controlled experiments are $a \in [-0.1, 0.1]$ and $b \in [0.1, 0.2]$. However, data can also be gathered "in the wild" by unobstrusively logging mouse cursor trajectories of everyday computer users [18], see Fig. 6 for an example. Compared

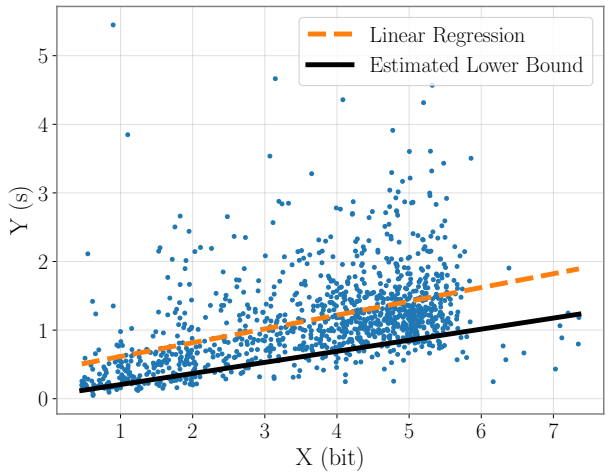

Fig. 6. Dataset of real-world aiming performance. Y is the time in seconds needed to select a target with difficulty X in bits.

to controlled data, "in the wild" data displays high variability and great positive skewness, because users do not routinely try to maximize their performance. Furthermore, technical difficulties associated with trajectory segmentation make some observations of Y appear as outliers. Linear regression yields parameters outside of the expected intervals ($a = 0.4$) or on the edge ($b = 0.2$), as well as a very low goodness of fit (coefficient of determination $r^2 = 0.2029$). We thus performed an EMG regression on the data of Fig. 6, with $\beta = \beta_0 + \beta_1\ x$ to take into account the fact that usually standard deviation scales with levels of X in psycho-physical experiments [10]. It was found that $a = 0.04$, $b = 0.16$, $\beta_0 = 0.33$, $\beta_1 = 0.05$, $\sigma^2 = 0.04$, bringing values of a and b inside trusted levels. The fitted lower bound is displayed Fig. 6 in black full line. Compared to linear regression, the lower bound estimated via EMG regression enables more meaningful comparison between controlled and "in the wild" experimental data.

VI. Conclusion and Future Work

This paper proposed a regression method to estimate a linear bound in the presence of outliers, based on an EMG noise model, which encompasses the classical linear regression and a minimum regression method as special cases. Simulation results have shown that the parameters a and b specifying the linear lower bound were consistently estimated, where the slope parameter b is unbiased and remarkably robust to the model mismatch. The EMG regression was illustrated successfully on a real-world example.

On more complicated examples where the lower bound appears not linear, it may be observed that the optimizer cannot find a global log-likelihood maximum. Future work will generalize the regression to more general models for μ (e.g., polynomial in x) and β. An interesting question is also whether the bias for a can be corrected, and whether the method is effective for very different parameter sets (e.g. slopes close to vertical $b = \infty$).

Acknowledgment

This research was partially funded by European Research Council (ERC) grant n^o 695464 ONE: Unified Principles of Interaction.

References

[1] C. Constable, "Parameter estimation in non-Gaussian noise," *Geophysical Journal International*, vol. 94, no. 1, pp. 131–142, 1988.

[2] V. K. Smith and T. W. Hall, "A comparison of maximum likelihood versus BLUE estimators," *The Review of Economics and Statistics*, pp. 186–190, 1972.

[3] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[4] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou, "Signal modeling and classification using a robust latent space model based on t-distributions," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 949–963, 2008.

[5] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[6] L. De Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. Springer Science & Business Media, 2007.

[7] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[10] R. D. Luce, *Response times: Their role in inferring elementary mental organization*. Oxford University Press on Demand, 1986, no. 8.

[11] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, and P. Tikhonov, "Reconstruction of chromatographic peaks using the exponentially modified Gaussian function," *Journal of Chemometrics*, vol. 25, no. 7, pp. 352–356, 2011.

[12] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *The Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[13] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust Region Methods*. Siam, 2000, vol. 1.

[14] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Vol. 1*. John Wiley & Sons, 1994.

[15] J. Gori, O. Rioul, and Y. Guiard, "Speed-accuracy tradeoff: A formal information-theoretic transmission scheme (FITTS)," *ACM Trans. Comput.-Hum. Interact.*, vol. 25, no. 5, pp. 27:1–27:33, Sep. 2018.

[16] J. Gori, O. Rioul, Y. Guiard, and M. Beaudouin-Lafon, "One Fitts' law, two metrics," in *IFIP Conference on Human-Computer Interaction*. Heidelberg, Germany: Springer, 2017, pp. 525–533.

[17] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement." *Journal of experimental psychology*, vol. 47, no. 6, p. 381, 1954.

[18] O. Chapuis, R. Blanch, and M. Beaudouin-Lafon, "Fitts' law in the wild: A field study of aimed movements," "LRI, article 1480, December 2007, 11 pages. [Online]. Available: http://insitu.lri.fr/ chapuis/publications/RR1480.pdf