MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1287 August 1991
C.B.I.P. Paper No. 66

Models of Noise and Robust Estimates

Federico Girosi

Abstract

Given n noisy observations g_i of the same quantity f, it is common use to give an estimate of f by minimizing the function $\sum_{i=1}^{n}(g_i - f)^2$. From a statistical point of view this corresponds to computing the Maximum Likelihood estimate, under the assumption of Gaussian noise. However, it is well known that this choice leads to results that are very sensitive to the presence of outliers in the data. For this reason it has been proposed to minimize functions of the form $\sum_{i=1}^{n} V(g_i - f)$, where V is a function that increases less rapidly than the square. Several choices for V have been proposed and successfully used to obtain "robust" estimates. In this paper we show that, for a class of functions V, using these robust estimators corresponds to assuming that data are corrupted by Gaussian noise whose variance fluctuates according to some given probability distribution, that uniquely determines the shape of V.

This paper describes research done within the Center for Biological Information Processing, in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Artificial Intelligence Center of Hughes Aircraft Corporation (S1-801534-2). Support for the A. I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010, and in part by ONR contract N00014-85-K-0124.

1 Introduction

A common problem in statistics is the following: given n noisy observations g_i of the same quantity f, give an estimate of f. A typical solution to this problem consists in choosing the value of f that maximes the *likelihood function* $P(g|f)$, that is the probability of having observed the data $g = (g_1, \ldots, g_n)$ if the true value was f. Estimates of this type are named Maximum Likelihood (ML) estimates, and rely on the assumption that we know the likelihood function $P(g_1, \ldots, g_n|f)$, that is essentially a model of how noise affected the measure process.

A common assumption is that of additive Gaussian noise, in which we assume that the measurement g_i are related to the true value by the relation

$$g_i = f + \epsilon_i \ , \quad i = 1, \ldots, n \ ,$$

where ϵ_i are independent random variables with given gaussian probability distributions $P_i(\epsilon_i)$ of variance σ_i and zero mean. In this case the likelihood function is

$$P(g_1, \ldots, g_n|f) = \prod_{i=1}^{n} P_i(\epsilon_i) = \prod_{i=1}^{n} \sqrt{\frac{\beta_i}{\pi}} e^{-\beta_i (g_i - f)^2} \tag{1}$$

where $\beta_i = \frac{1}{2\sigma_i^2}$. Maximizing the likelihood function (1) corresponds therefore to solve the following minimization problem:

$$\min_f \sum_{i=1}^{N} \beta_i (g_i - f)^2 \ . \tag{2}$$

An elementary computation shows that the solution is the weighted average of the data:

$$f = \frac{\sum_{i=1}^{n} \beta_i g_i}{\sum_{i=1}^{n} \beta_i} \ .$$

The ML estimate has therefore a simple meaning and it is easy to compute. However, it is well known that estimates of this type are not "robust", that is are they very sensitive to the presence of outliers in the data. In order to overcome this difficulty it has been proposed to use a modified version of the minimization problem (2):

$$\min_f \sum_{i=1}^{N} V(g_i - f) \ , \tag{3}$$

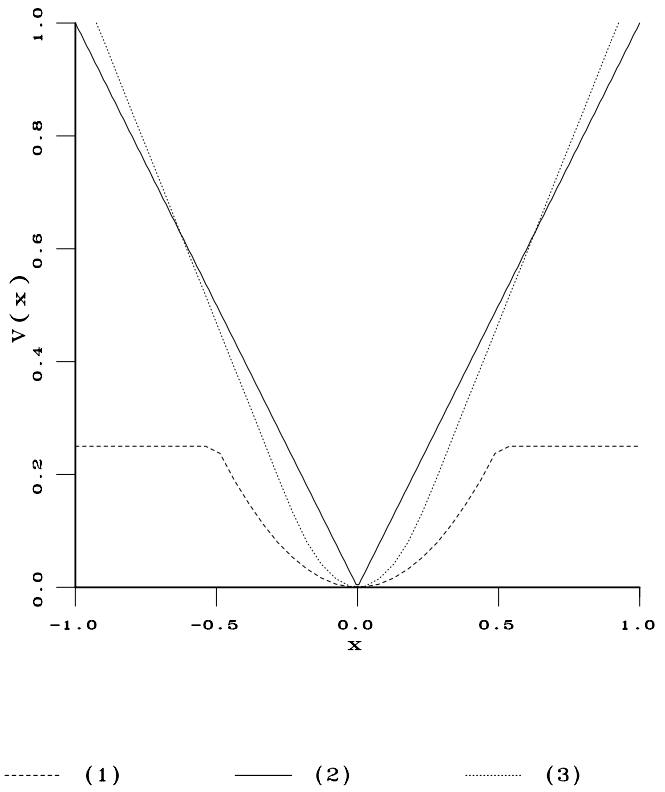

Figure 1: Different choices for the function V. (1) $V(x)$ is quadratic for $x < 0.5$ and then constant. (2) $V(x) = |x|$. (3) $V(x)$ is quadratic for $x < 0.5$ and then linear, with continuous first derivative.

where the quadratic function $(g_i - f)^2$ has been substituted by some other less rapidly increasing even function V. Estimator of this type are known in statistics, for particular choices of V, as *robust estimators* (Huber, 1981). The idea underlying (3) is that if the error $(g_i - f)^2$ is large, it is likely that g_i is an outlier, so that we do not want to enforce f to be close to it. Therefore the function V should not increase much after a certain value. Different shapes for V have been proposed, and some of them have been depicted in figure (1).

In this paper we want to give a more rigorous justification for the use of estimates like the one of eq. (3), and also to give an interpretation of the model of noise to which they correspond. We will see that if the function $e^{-V(x)}$ is completely monotone, then using eq. (3) corresponds to assuming that our measures are affected by a Gaussian noise whose variance is a random variable with given probability distribution. Depending on the probability distribution of the variance of the noise, different shapes for V are obtained. For a particular choice of V a justification of such a technique was given in (Girosi, Poggio and Caprile, 1991), but no characterization was given. In the next section we formalize these statements, while in the following sections we present a large class of functions V that can be used, together with some

examples.

2 Robust Maximum Likelihood Estimates

In order to simplify the notation we consider the problem presented in the previous section in which only one measurement g is done, since this does not change the main conclusions. We therefore assume that

$$g = f + \epsilon \tag{4}$$

where ϵ is a random variable whose distribution is Gaussian with zero mean and variance σ. The likelihood function is therefore

$$P(g|f) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(g-f)^2} \tag{5}$$

where $\beta = \frac{1}{\sigma^2}$.

When we compute the standard maximum likelihood estimate we are assuming that the variance of the noise has a fixed value, but this assumption is not always realistic. In fact, in many cases the accuracy of the measurement apparatus can fluctuate, due to some external causes, and in these cases our data can contain outliers. A more realistic assumption consists in considering the variance of the Gaussian noise, and therefore β, as a random variable, with given distribution $P(\beta)$. We are therefore led to introduce the probability $P(g|f,\beta)$ of having observed the data g if the true value was f and the variance of the noise was $\sigma = \frac{1}{\sqrt{\beta}}$:

$$P(g|f,\beta) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(g-f)^2} \ . \tag{6}$$

Notice that the right hand side of eq. (6) is the same of eq. (5), but their meaning is different. We can now compute the joint probability $P(g,\beta|f)$ of having observed the data g in presence of gaussian noise with variance $\sigma = \frac{1}{\sqrt{\beta}}$, if the true value was f:

$$P(g,\beta|f) = P(g|f,\beta)\ P(\beta)\ . \tag{7}$$

Since we are not interested in estimating β, but we are interested only in the probability of g given f, that is our likelihood function, we integrate equation (7) over β to obtain the *effective noise distribution*

$$P^*(g|f) = \int_0^\infty P(g|f,\beta)\ P(\beta)d\beta = \frac{1}{\sqrt{\pi}} \int_0^\infty e^{-\beta(g-f)^2} \sqrt{\beta}\ P(\beta)d\beta \qquad (8)$$

The MAP estimate is now obtained by maximizing the probability of eq. (8), or, taking the negative of its logarithm, solving the following minimization problem:

$$\min_f\ V(g-f) \qquad (9)$$

where we have defined the so called *effective potential*[1]

$$V(x) = -\ln \int_0^\infty e^{-\beta x^2} \sqrt{\beta}\ P(\beta)d\beta\ . \qquad (10)$$

In the case in which n observations $g_1, \ldots, g_n$ have been taken the same considerations apply, and assuming that the variances σ_i of the measurements g_i have all the same probability distribution, we obtain, instead of eq. (9):

$$\min_f\ \sum_{i=1}^{N} V(g_i - f)\ . \qquad (11)$$

This equation coincides with eq. (3), that has been proposed has a technique to "robustize" the least-square estimate (2). In our case, however, the effective potential V derives from specific assumptions on how data are corrupted by noise. If the distribution of the random variable β is a delta function centered on some value $\bar{\beta}$, that is if $P(\beta) = \delta(\beta - \bar{\beta})$, the noise model is Gaussian with fixed variance, and the effective potential is a quadratic function, yielding the same result of eq. (2). For other probability distributions $P(\beta)$, formula (10) allows to compute the corresponding effective potential by simply performing a one dimensional integration. Conversely, in some cases, given an effective potential $V(x)$, it is also possible to understand if there is any probability distribution $P(\beta)$ that corresponds to it. In the next section we introduce a class of effective noise distributions for which such a characterization can be given.

3 A class of effective noise distributions

In this section we study and characterize a class of effective noise distributions. Since we want to maximize the effective noise distribution (8) we

[1]This name was previously introduced by Geiger and Girosi (1991), that used a similar technique applied at the problem of surface reconstruction with discontinuities.

are not interested in effective distributions that are unbounded. It will turn out that if an effective noise distribution is bounded at the origin it is also bounded on all the real axis. Therefore, according to eq. (8) we define the *bounded effective noise distributions* as the probability distributions of the form:

$$f(x) = \int_0^\infty e^{-\beta x^2} \sqrt{\beta} P(\beta) d\beta \tag{12}$$

where $P(\beta)$ is a probability distribution, and such that the following condition is satisfied:

$$f(0) = \int_0^\infty \sqrt{\beta} P(\beta) d\beta < +\infty \ .$$

We can now prove the following proposition:

Proposition 3.1 *A probability distribution $f(x)$ is a bounded effective noise distributions if and only if $f(\sqrt{x})$ is completely monotone.*

Proof: (*only if*) Suppose $f(x)$ is a bounded effective noise distribution. Then $f(\sqrt{x})$ it can be represented as

$$f(\sqrt{x}) = \int_0^\infty e^{-\beta x} d\mu(\beta)$$

where

$$\mu(\beta) = \int_0^\beta \sqrt{\tau} P(\tau) d\tau \ .$$

Since $\mu(\beta)$ is clearly non decreasing and bounded, then by the Bernstein's theorem on the representation of completely monotone functions (see Appendix A), $f(\sqrt{x})$ is completely monotone.

(*if*) Suppose that the probability distribution $f(x)$ is such that $f(\sqrt{x})$ is completely monotone. Then it can be represented as

$$f(x) = \int_0^\infty e^{-\beta x^2} d\mu(\beta) \ , \tag{13}$$

with $\mu(\beta)$ non decreasing and bounded. Since f is a probability distribution its integral over the real axis has unit value, and therefore

$$1 = \int_{-\infty}^{+\infty} f(x) = 2 \int_0^\infty \left(\int_0^\infty e^{-\beta x^2} d\mu(\beta) \right) dx$$

Exchanging the order of integration and evaluating the gaussian integral, we obtain that

$$\int_0^\infty \frac{d\mu(\beta)}{\sqrt{\beta}} = c \ , \quad 0 < c < +\infty \ .$$

Therefore it is always possible to write

$$d\mu(\beta) = P(\beta) \ \sqrt{\beta} \ d\beta \ ,$$

where $P(\beta)$ is a probability distribution, being positive and having finite integral. Substituting this expression in formula (13) we obtain the representation of eq. (12). Noticing that completely monotone functions are bounded at the origin, since

$$f(0) = \int_0^\infty d\mu(\beta) < +\infty \ ,$$

we conclude that $f(x)$ is a bounded effective noise distribution.□

We can now answer to the question if effective potentials of the type $V(x) = |x|^p$ can be derived in this framework. In fact, using the previous proposition it is sufficient to check if the probability distribution $P(x) = e^{-|x|^p}$ is such that $P(\sqrt{x})$ is completely monotone. Using the fact that the function $e^{-|x|^p}$ is completely monotone if and only if $0 < p \leq 1$ (Schoenberg, 1937)(see appendix A), we can immediately derive the following proposition:

Proposition 3.2 *The function $V(x) = |x|^p$ is the effective potential associated to a bounded effective noise distribution if and only if $0 < p \leq 2$.*

We notice that if we set $p = 1$ in the proposition above we obtain as effective potential the usual L_1 error measure, that is $V(x) = |x|$, is obtained. However, since the function absolute value is not differentiable at the origin it has been proposed to use functions that behave quadratically in a neighbor of the origin, and linearly for large values of the argument (Eubank, 1988). Effective potentials of the form $V(x) = |x|^p$ are interesting, since they are convex and the problem of maximizing the likelihood function has therefore only one solution. However, before showing what are the effective noise distributions that are associated to this effective potentials, we present a more simple example, that gives a non convex effective potential that has also been used in practice.

4 A class of non convex effective potentials

We have already seen that if the distribution $P(\beta)$ is a delta function the standard quadratic potential is obtained. The simplest non trivial case consists in assuming that $P(\beta)$ is a sum of two delta functions, that is

$$P(\beta) = (1-\epsilon)\delta(\beta-\beta_1) + \epsilon\delta(\beta-\beta_2) \tag{14}$$

where ϵ is a parameter between 0 and 1 and β_1, β_2 are fixed positive numbers, $\beta_1 > \beta_2$. If β_2 is a very small number such a distribution can represent the a priori knowledge that a fraction ϵ of the data is very unreliable. In the limit of β_2 going to zero this fraction of data is constituted by genuine outliers, and we therefore analyze the model keeping in mind that we are interested in this limit.

With the noise distribution given by eq. (14) the effective potential becomes

$$V(x) = -\ln \int_0^\infty \sqrt{\beta} e^{-\beta x^2} [(1-\epsilon)\delta(\beta-\beta_1) + \epsilon\delta(\beta-\beta_2)] d\beta \tag{15}$$

and, after some algebra:

$$V(x) = \beta_1 x^2 - \ln\left(1 + \frac{\epsilon}{1-\epsilon}\sqrt{\frac{\beta_2}{\beta_1}} e^{x^2(\beta_1-\beta_2)}\right) \ , \tag{16}$$

where we have neglected unimportant constant terms.

We start studying the behavior of the potential in a neighbor of the origin. Taking a Taylor's expansion up to the second order, after some algebra we find that

$$V(x) = V(0) + \beta_2 x^2 + o(x^3)$$

so that the potential is initially quadratic, and very flat if β_2 is small, that is if we assume that outliers are present in the data.

When x goes to infinity the exponential term in the logarithm of eq. (16) grows very fast, (remember that $\beta_1 > \beta_2$), and the unit term can be omitted, leading to major simplifications. This is true only if β_2 is "not to small", in the sense that the following inequality has to verified:

$$x^2 >> k(\epsilon, \beta_1) - \frac{1}{2}\ln\beta_2 \tag{17}$$

where $k(\epsilon, \beta_1)$ is a constant that depends only ϵ and β_1, whose exact form is irrelevant to us. In the region where this condition is satisfied we therefore obtain:

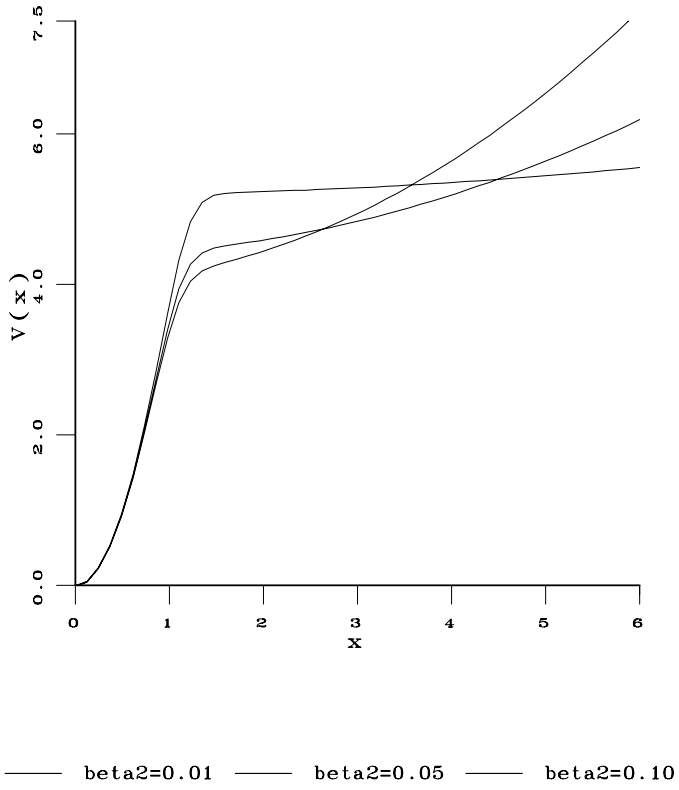

Figure 2: The non convex effective potential of the model above for different values of β_2.

$$V(x) \approx \beta_1 x^2 - \ln k(\epsilon, \beta_1) - \frac{1}{2}\ln \beta_2 - x^2(\beta_1 - \beta_2)$$

and therefore:

$$V(x) \approx \beta_2 x^2 - \ln k(\epsilon, \beta_1) - \frac{1}{2}\ln \beta_2 \ .$$

For large values of x and small values of β_2, where "large" and "small" have to be intended in the sense of condition (17), the effective potential is again quadratic and very flat.

In summary: for small values of x the potential is a very flat parabola, for large values of x is the same parabola, but translated of a positive amount that grows logarithmically with β_2, and in between, since its first derivative is strictly positive, it smoothly connects these two behaviors. In fig. (2) we show the shape of the effective potential for fixed ϵ and β_1, for three different values of β_2. We set $\epsilon = 0.1$, $\beta_1 = 4$, and $\beta_2 \in \{0.1,\ 0.05,\ 0.01\}$. This amounts to say that we know a priori that 90% of the data points are affected by Gaussian noise of variance equal to 0.5 (that is $\sqrt{\frac{1}{4}}$). The other 10% is affected by Gaussian noise with very large variance, that is $\sigma = 3.16,\ 4.47,\ 10$. We notice that for a value of $\beta_2 = 0.01$, that corresponds to a variance

$\sigma = 10$, the effective potential is extremely flat, almost constant. A similar behavior is expected: in fact it means that when the interpolation error is larger than a threshold its influence on the solution is not taken in account anymore, and this is exactly the kind of motivation that led statisticians to consider *robust* models.

5 A class of convex effective potentials

We now consider an effective potential of the form

$$V(x) = \sqrt{\alpha^2 + x^2} \ . \tag{18}$$

where α is some given parameter, possibly zero. Functions of this type are well known in approximation theory by the name of "multiquadrics", or "Hardy's multiquadrics", and their behavior is shown in fig. (3). Potentials of this shape are interesting because they are convex, so that the minimization problem associated with them has a unique solution. Moreover, potentials with a shape very similar to this one can be implemented in analog VLSI circuits (Harris, 1990), allowing very fast ways to solve the estimation problems.

We are interested in finding the probability distribution that leads to this form of effective potential. A solution to this problem certainly exists, since it is easy to show that $e^{-V(\sqrt{x})}$ is completely monotonic. Therefore we have to find a function $P(\beta)$ such that

$$e^{-\sqrt{\alpha^2+x^2}} = \int_0^\infty e^{-x^2\beta} \sqrt{(}\beta) P(\beta) \ d\beta,$$

This is in essence the problem of computing an inverse Laplace transform. We start from the following identity (Gradshteyn and Ryzhik, 1981):

$$2\sqrt{\pi} e^{-\sqrt{x}} = \int_0^\infty \beta^{-\frac{3}{2}} e^{-\frac{1}{4\beta}} e^{-x\beta} \ d\beta \tag{19}$$

and perform the substitution $x \to x^2 + \alpha^2$, obtaining

$$2\sqrt{\pi} e^{-\sqrt{x^2+\alpha^2}} = \int_0^\infty \beta^{-\frac{3}{2}} e^{-\frac{1}{4\beta}} e^{-\beta x^2 - \beta\alpha^2} \ d\beta \ . \tag{20}$$

Making the proper identifications in equation above, and paying attention to normalization factors, we obtain as a result:

$$P(\beta) = \frac{1}{\beta^2} e^{-\frac{1}{4\beta} - \beta\alpha^2 + \alpha} \ . \tag{21}$$

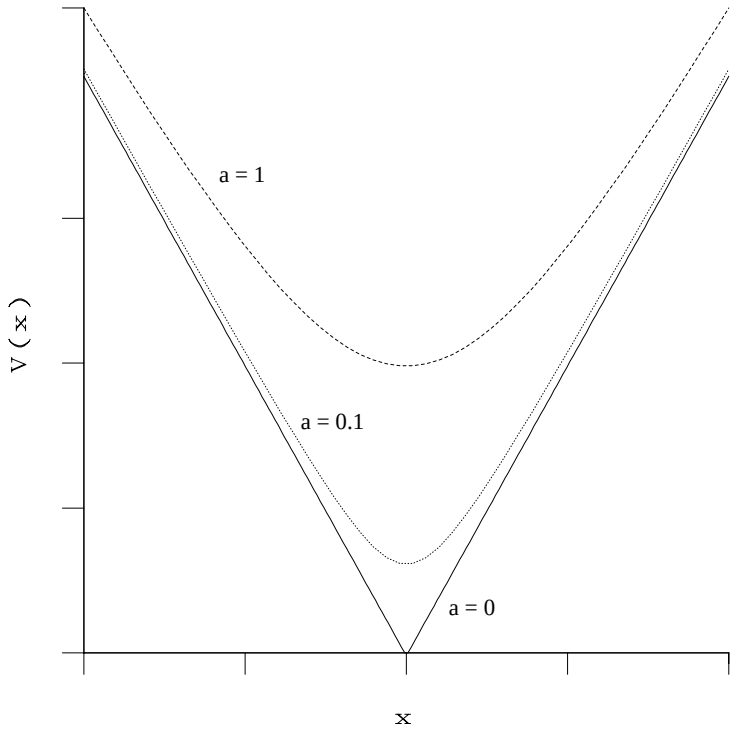

Figure 3: The multiquadric effective potential $V(x) = (a^2 + x^2)^{\frac{1}{2}}$ for three different values of the parameter a.

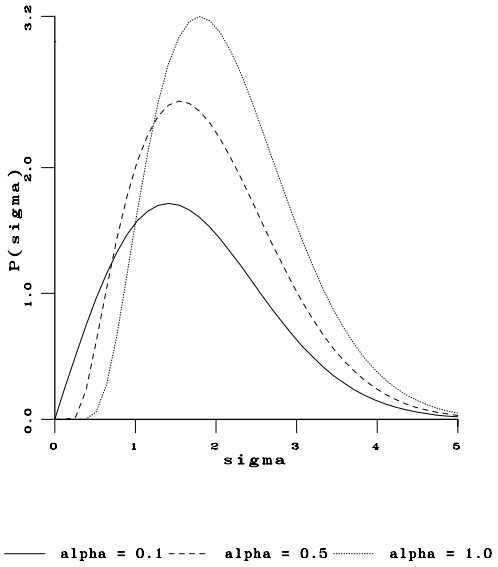

Figure 4: The distribution of variance $\tilde{P}(\sigma)$ associated to the multiquadric effective potential.

We can now derive the distribution $\tilde{P}(\sigma)$ of the variance $\sigma = \frac{1}{\sqrt{\beta}}$, imposing

$$P(\beta)d\beta = \tilde{P}(\sigma)d\sigma \ , \qquad \beta = \frac{1}{\sigma^2} \ . \tag{22}$$

After some algebra we obtain the function

$$\tilde{P}(\sigma) = 2\sigma e^{-\frac{\sigma^2}{4} - \frac{\alpha^2}{\sigma^2} + \alpha} \ .$$

whose shape is depicted in fig, (4) for three different values of α. We notice that when α increases the distribution becomes more peaked, and also flatter around the origin. Therefore the probability of having low-noise data decreases when α increases. Equivalently, we can also say that the probability of having data with noise larger than a given threshold increases with α.

6 Conclusions

We have shown that it is possible to give a simple interpretation to estimators based on the solution of the minimization problem

$$\min_f \sum_{i=1}^{N} V(g_i - f) \ , \tag{23}$$

where V is an appropriate function, that we call effective potential. If the function $e^{-V(\sqrt{x})}$ is completely monotone, using these robust estimators corresponds to compute Maximum Likelihood estimators under the assumption that data are corrupted by Gaussian noise whose variance fluctuates according to a given probability distribution, that uniquely determines V. Typical "effective potentials" V, that have been used in the past, belongs to the class we consider.

We notice that the result we derived holds also in the more general settings context of parametric and non parametric regression. In order to see why, let $g = \{(\mathbf{x}_i, y_i) \in R^n \times R\}_{i=1}^N$ be a set of data that has been obtained by randomly sampling a multivariate function f in presence of noise. In parametric regression we assume that f is a parametric function $h(\mathbf{x}; \mathbf{p})$, where $\mathbf{p} \in R^m$, and the optimal set of parameters $\mathbf{p}$ is usually recovered by minimizing the least square error

$$\min_{\mathbf{p} \in R^m} \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i; \mathbf{p}))^2 \ . \tag{24}$$

As in the case considered in this paper, this can be thought as a maximum likelihood estimator, under the assumption of Gaussian noise with fixed variance. Therefore the same argument we applied in section (2) applies here, and more robust estimates could be obtained if we replace the quadratic function in eq. (24) with an effective potential V.

In non parametric regression no assumption is made on the specific form of f, and a common technique consists in solving the following minimization problem:

$$\min_f \ \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2 + \lambda S[f] \ .$$

where $S[f]$ is an appropriate convex functional of f and λ a positive number. This correspond to compute the Maximum A Posteriori estimator, under the assumption of Gaussian noise and a priori probability for f given by $P(f) \propto e^{-\lambda S[f]}$. If we assume that the variance of the Gaussian noise is a random variable, using the same argument we used in section (2) we can prove that the Maximum A Posteriori estimator solves the following minimization problem:

$$\min_f \ \sum_{i=1}^{N} V(f(\mathbf{x}_i) - y_i) + \lambda S[f] \ . \tag{25}$$

where V is an effective potential. Estimators of this type are known in the statistical literature as *M-type smoothing splines* (Eubank, 1988), and their implementation in analog VLSI circuits has been considered by J. Harris (1991) for some choices of the functional $S[f]$.

A Completely Monotone Functions

We need to give the following:

Definition A.1 *A function f is said to be completely monotonic on $(0,\infty)$ provided that it is $C^\infty(0,\infty)$ and $(-1)^l \frac{\partial^l f}{\partial x^l}(x) \geq 0$, $\forall x \in (0,\infty)$, $\forall l \in \mathcal{N}$, where $\mathcal{N}$ is the set of natural numbers.*

A typical example of completely monotone function is the exponential function $f(x) = e^{-\alpha x}$, with $\alpha > 0$. It turns out that all the completely monotone functions are linear superpositions with positive coefficients of scaled exponentials, as the following theorem of Bernstein shows:

Theorem A.1 (Bernstein, 1929) *The class of completely monotone functions is identical with the class of functions of the form*

$$g(x) = \int_0^\infty e^{-\beta x}\, d\mu(\beta),$$

where $\mu(\beta)$ is non-decreasing and bounded for $\beta \geq 0$.

References

[1] S. Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52:1–66, 1929.

[2] R.L. Eubank. *Spline Smoothing and Nonparametric Regression*, volume 90 of *Statistics, textbooks and monographs*. Marcel Dekker, Basel, 1988.

[3] D. Geiger and F. Girosi. Parallel and deterministic algorithms for MRFs: surface reconstruction and integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):401–412, May 1991.

[4] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximation and learning: outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[5] I.S. Gradshteyn and I.M. Ryzhik. *Table of integrals, series, and products.* Academic Press, New York, 1981.

[6] J.G. Harris, S. Liu, and B. Mathur. Discarding outliers in a nonlinear resistive network. In *International Joint Conference on Neural Networks*, Seattle, WA., to appear., July 1991.

[7] P.J. Huber. *Robust Statistics.* John Wiley and Sons, New York, 1981.

[8] I.J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math.*, 39:811–841, 1938.