Journal of Uncertain Systems
Vol.5, No.2, pp.137-140, 2011
Online at: www.jus.org.uk

Power Law Regularization in Probabilistic Inverse Problems: a Theoretical Justification

Kanita Petcharat, Piyaphon Paichit, Sa-aat Niwitpong*
Department of Applied Statistics, Faculty of Applied Science
King Mongkut's University of Technology North Bangkok,
1518 Pibulsongkram Rd., Bangsue, Bangkok 10800 Thailand

Received 9 January, 2009; Revised 30 January, 2010

Abstract

In many practical situations, it is desirable to reconstruct the probability distribution $\rho(x)$ of a quantity x from the noisy measurements of this quantity. This reconstruction often results in heavier tails, so to get a more accurate reconstruction, it is reasonable to decrease these tails. An empirical comparison of several possible tail-decreasing strategies has shown that the best possible strategy is to raise all the values of the reconstructed density $\widetilde{\rho}(x)$ to a power $a > 1$, and then normalize the resulting distribution.

In this paper, we provide a theoretical explanation for this empirical fact.

Keywords: inverse problem, power law regularization, theoretical justification

1 Formulation of the Problem

Reconstructing probability distributions is a practically important problem. In many practical situations, we need to find the probability density $\rho(x)$ of a physical quantity x.

Reconstructing probability distributions: traditional statistical approach. Traditional statistical approach assumes that we have a sample of values $x_1, \ldots, x_n$ distributed according to this distribution. Based on this sample, we can then reconstruct the desired probability density function $\rho(x)$; see, e.g., [4]

In practice, we must take measurement errors into account. In practice, the values x_i come from measurement, and measurement is never absolutely accurate: what we actually observe are values $m_i = x_i + e_i \neq x_i$, where $e_i = m_i - x_i$ are measurement errors.

Thus, what we reconstruct by using the traditional statistical techniques is *not* the probability density for the desired quantity x, but rather the probability density $\rho_m(x)$ of the sum $m = x + e$.

How measurement errors can be taken into account. Measurement errors are usually independent from the desired quantity, so it is reasonable to assume that the corresponding random variables x and e are independent. Under this assumption, the observed probability density $\rho_m(x)$ is related to the probability densities $\rho(x)$ and $\rho_e(x)$ by the known convolution relationship:

$$\rho_m(x) = \int \rho(y) \cdot \rho_e(x - y)\, dy. \tag{1}$$

By performing measurements in the absence of the signal (or for a well-calibrated signal), we can determine the probability distribution $\rho_e(x)$ of the measurement errors. Thus, we know $\rho_m(x)$ and $\rho_e(x)$ and so, in principle, we can solve the equation (1) and reconstruct the desired probability distribution $\rho(x)$; see, e.g., [2] and references therein.

*Corresponding author. Email: snw@kmutnb.ac.th (S. Niwitpong).

Need for further improvement. Because of the noise (= measurement errors), the reconstructed distribution $\widetilde{\rho}(x)$ is somewhat different from the actual distribution $\rho(x)$. It is therefore desirable to decrease this difference.

How to improve the reconstructed distribution: analysis of the problem. The difference between the actual and the reconstructed distributions is especially large for the values x which occur rarely – and for which, therefore, there are few data points.

In other words, the reconstructed values $\widetilde{\rho}(x)$ may be reasonably accurate when $\rho(x)$ is large, but the values $\widetilde{\rho}(x)$ may be corrupted when $\rho(x)$ is small – i.e., at the tails of the original distribution.

When the actual values $\rho(x)$ are close to 0, the only drastic difference is when the reconstructed values $\widetilde{\rho}(x)$ are much larger. So, due to the measurement errors, the tails of the reconstructed distribution are heavier than of the actual one.

Thus, to improve the reconstructed distribution $\widetilde{\rho}(x)$, it is desirable to decrease its tails.

How to decrease the tails: a general idea. A reasonable idea is to select an appropriate tail-decreasing function $f(x)$ and then replace each value $\widetilde{\rho}(x)$ with a new value $f(\widetilde{\rho}(x))$ in such a way that when $\widetilde{\rho}(x)$ is small, the new value $f(\widetilde{\rho}(x))$ is even smaller.

This replacement may change the overall probability value from

$$\int \widetilde{\rho}(x)\,dx = 1 \tag{2}$$

to

$$\int f(\widetilde{\rho}(x))\,dx \neq 1. \tag{3}$$

So, to get a probability density function, we must normalize the values $f(\widetilde{\rho}(x))$, i.e., take the new probability density function

$$\rho_c(x) = \frac{f(\widetilde{\rho}(x))}{\int f(\widetilde{\rho}(y))\,dy}. \tag{4}$$

Empirical fact: power-law functions are the best for tail decrease. An empirical analysis described in [2] showed that the most adequate reconstruction occurs when we use power-law functions $f(x) = x^a$ for some $a > 1$.

What we do in this paper. In this paper, we provide a theoretical explanation for the above empirical fact.

2 A New Theoretical Justification

The main idea behind our explanation. The numerical value of a physical quantity depends on the choice of a measuring unit. For example, 1 inch is exactly the same distance as 2.54 cm. It is natural to require that the results of the tail decrease procedure not change if we simply change the measuring unit. We will show that this reasonable requirement uniquely determines the power law functions.

Towards formalizing the main idea. If we change the original measuring unit to a new one which is λ times larger, then the quantity that is described by the value x in the old units will have a numerical value $x' = x/\lambda$ in the new units.

The probability density $\widetilde{\rho}(x)$ means that the probability to find the value between x and $x + dx$ is equal to $\widetilde{\rho}(x) \cdot dx$. In terms of the new units $x' = x/\lambda$, this same value $\widetilde{\rho}(x) \cdot dx$ is the probability to find the value x' between $x' = x/\lambda$ and $x' + dx'$, where $dx' = dx/\lambda$. From $x' = x/\lambda$, we conclude that $x = x' \cdot \lambda$ and $dx = dx' \cdot \lambda$. Thus, the probability to find a value between x' and $x' + dx'$ is equal to

$$\widetilde{\rho}(x) \cdot dx = \widetilde{\rho}(\lambda \cdot x') \cdot \lambda \cdot dx'. \tag{5}$$

On the other hand, by definition of the probability density $\widetilde{\rho}'(x')$, this same probability is equal to

$$\widetilde{\rho}'(x) \cdot dx'. \tag{6}$$

By equating the expressions (5) and (6), we conclude that

$$\widetilde{\rho}'(x') = \widetilde{\rho}(\lambda \cdot x') \cdot \lambda. \tag{7}$$

From the main idea to the justification of the power law. If we apply the tail decrease procedure with a function $f(x)$ to the probability density function $\widetilde{\rho}(x)$ described in the original units, we get a new function

$$\rho_c(x) = c \cdot f(\widetilde{\rho}(x)) \tag{8}$$

for an appropriate normalizing constant c. In terms of the new units x', this new probability density function has the form

$$\rho_c'(x') = \rho_c(\lambda \cdot x') \cdot \lambda = c \cdot f(\widetilde{\rho}(\lambda \cdot x')). \tag{9}$$

On the other hand, if we directly apply the same tail decrease procedure to the expression (7) describing the probability density in the new units, we get an expression

$$\rho_c''(x') = c' \cdot f(\widetilde{\rho}'(x')) \tag{10}$$

for an appropriate normalizing constant c'. Substituting the expression (7) for $\rho'(x')$ into this formula, we conclude that

$$\rho_c''(x') = c' \cdot f(\widetilde{\rho}(\lambda \cdot x') \cdot \lambda). \tag{11}$$

Our requirement is that the result of tail decrease should not depend on whether we apply this procedure in the original units or in the new units, i.e., that the resulting probability density functions $\rho_c'(x')$ and $\rho_c''(x')$ must coincide. By equating the right-hand sides of the expressions (9) and (11), we get

$$c' \cdot f(\widetilde{\rho}(\lambda \cdot x') \cdot \lambda) = c \cdot f(\widetilde{\rho}(\lambda \cdot x')). \tag{12}$$

This must be true for all values $\widetilde{\rho}$, so we conclude that

$$c' \cdot f(\widetilde{\rho} \cdot \lambda) = c \cdot f(\widetilde{\rho}). \tag{13}$$

Dividing both sides of this equality by c' and taking into account that the ratio $r \stackrel{\text{def}}{=} c/c'$ may depend on λ, we deduce that for every two numbers $\widetilde{\rho} > 0$ and λ, the following equality holds:

$$f(\widetilde{\rho} \cdot \lambda) = r(\lambda) \cdot f(\widetilde{\rho}). \tag{14}$$

It is known that every continuous function $f(x)$ satisfying this property has the form $f(x) = C \cdot x^a$ for some a; see, e.g., [1], Section 3.1.1. (This result was first proven in [3].)

The coefficient C does not matter since we normalize the probability density function anyway, so we can simply conclude that $f(x) = x^a$.

Conclusion. The only tail decrease procedures for which the results do not depend on the choice of the measuring units are the power-law procedures, with $f(x) = x^a$.

Comment. For differentiable functions $f(x)$, the result abut power functions is easy to prove. Indeed, if we differentiate both sides of (14) by λ and take $\lambda = 1$, we get

$$\widetilde{\rho} \cdot \frac{df}{d\widetilde{\rho}} = a \cdot f, \tag{15}$$

where $a \stackrel{\text{def}}{=} r(1)$. By moving all the terms containing f into one side and all the terms containing $\widetilde{\rho}$ to the other side, we conclude that

$$\frac{df}{f} = a \cdot \frac{d\widetilde{\rho}}{\widetilde{\rho}}. \tag{16}$$

Integrating both sides, we get

$$\ln(f) = a \cdot \ln(\widetilde{\rho}) + c, \tag{17}$$

hence

$$f = e^{a \cdot \ln(\widetilde{\rho}) + c} = e^c \cdot \left(e^{\ln(\widetilde{\rho})}\right)^a = C \cdot \widetilde{\rho}^a \tag{18}$$

for $C = e^c$.

3 Observation: Tail Decrease Procedure is Related to the Notion of Entropy

When the measurements are precise, we get the exact values of the probabilities, i.e., we should take $f(x) = x$, with $a = 1$. When the measurements are very accurate, we get the probabilities close to the actual ones, so the tail decrease procedure $f(x) = x^a$ should minimally change the probability values – and thus, we should have $a = 1 + \varepsilon$ for some small ε.

For a small ε, we can expand the expression $\rho^{1+\varepsilon}$ in Taylor series in ε and keep only linear terms in this expansion (thus ignoring quadratic and higher order terms). As a result, we get

$$\widetilde{\rho}^{1+\varepsilon} = \widetilde{\rho} + \varepsilon \cdot \widetilde{\rho} \cdot \ln(\widetilde{\rho}). \tag{19}$$

Indeed,

$$\widetilde{\rho}^{1+\varepsilon} = \widetilde{\rho} \cdot \widetilde{\rho}^{\varepsilon} = \widetilde{\rho} \cdot (e^{\ln(\widetilde{\rho})})^{\varepsilon} = \widetilde{\rho} \cdot e^{\varepsilon \cdot \ln(\widetilde{\rho})} \approx \widetilde{\rho} \cdot (1 + \varepsilon \cdot \ln(\widetilde{\rho})). \tag{20}$$

Due to (19), the normalizing constant is equal to

$$\int f(\widetilde{\rho}(y))\, dy = \int \widetilde{\rho}(y)\, dy + \varepsilon \cdot \int \widetilde{\rho}(y) \cdot \ln(\widetilde{\rho}(y))\, dy. \tag{21}$$

The first integral is the full probability, i.e., 1, the second integral differs only by the sign from the entropy

$$S(\widetilde{\rho}) = -\int \widetilde{\rho}(y) \cdot \ln(\widetilde{\rho}(y))\, dy \tag{22}$$

of the probability distribution, thus,

$$\int f(\widetilde{\rho}(y))\, dy = 1 - \varepsilon \cdot S(\widetilde{\rho}), \tag{23}$$

and the resulting normalized distribution (4) takes the form

$$\rho_c(x) = \frac{(\widetilde{\rho}(x))^{1+\varepsilon}}{1 - \varepsilon \cdot S(\widetilde{\rho})}. \tag{24}$$

Acknowledgments

The authors are thankful to Vladik Kreinovich for his help and encouragement.

References

[1] Aczel, J., *Lectures on Functional Equations and Their Applications*, Dover Publ., New York, 2006.

[2] Ferson, S., Using approximate deconvolution to estimate cleanup targets in probabilistic risk analyses, *Hydrocarbon Contaminated Soils*, edited by P.T. Kostecki, E.J. Calabrese and M. Bonazountas, Amherst Scientific Publishers, Amherst, Massachesetts, vol.5, pp.245–254, 1995.

[3] Pexider, J., Notiz uber Funktionaltheoreme, Monatsch. Math. Phys., vol.14, pp.293–301, 1903.

[4] Sheskin, D.J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC Press, Boca Raton, Florida, 2007.