

Robust Non-Parametric Curve Estimation using Density Power Divergences

Arun Kumar Kuchibhotla, Prof. Ayanendranath Basu

16 January 2015



M-estimation in Linear Regression

- M-estimation as was brought into limelight by Huber (1964) has been extensively studied in location (and location-scale) model.
- Similarly, M-estimation of multiple linear regression model has been extensively studied. For example, LTS, LMS, MM-estimators, S-estimators etc. Linear regression model is given by

$$y_i = x_i^\top \beta_0 + \epsilon_i, \quad \epsilon_1, \epsilon_2, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, 1).$$

- Most of these estimators generalize the normal likelihood estimating equation (also least squares).
- **MLE**: minimize over β , $\frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \beta)^2$.
- **M-estimator**: minimize over β , $\frac{1}{n} \sum_{i=1}^n \phi(y_i - x_i^\top \beta)$ for some suitably chosen function ϕ .

Likelihood in NPs

- Consider the non-parametric regression model:

$$y_i = m_0(x_i) + \epsilon_i, \quad \epsilon_1, \epsilon_2, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, 1).$$

- In order to avoid over-fitting, there are two proposals available in literature.
- P1**: minimize over $f \in C^2$,

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \int_{[0,1]^d} f''(x)^2 dx.$$

- P2**: minimize over $\theta \in \mathbb{R}$ for each $x \in [0, 1]^d$,

$$\frac{1}{n} \sum_{i=1}^n W_i(x)(y_i - \theta)^2,$$

to get $\hat{m}(x)$. Here $W_i(x)$ denotes the weights. For example

$$W_i(x) = \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right) \text{ for a kernel } K.$$

M-estimation in NPs

- Analogs of these two procedures were studied in literature in the context of M-estimation to get robust estimates in non-parametric regression.
- M-estimation similar to **P1** was studied by Huber (1979), Utreras (1981), Cox (1983), Chaudhuri (1995) (from the likelihood point of view).
- M-estimation similar to **P2** was studied by Cleveland (1979); Härdle (1984); Boente and Fraiman (1989).
- Extensions of these methods with simultaneous scale estimation are also available. See for example Härdle and Tsybakov (1988).
- Most of these works study non-parametric estimation without considering a model for errors (similar to M-estimation of location).

Density Power Divergences

- Density power divergences as was proposed by Basu et al. (1998) offers a smooth extension of maximum likelihood method of estimation.

$$\rho_{\alpha}(g, f) = \frac{1}{\alpha} \int g^{1+\alpha}(x) dx - \frac{1+\alpha}{\alpha} \int g(x) f^{\alpha}(x) dx + \int f^{1+\alpha}(x) dx.$$

- Suppose we now have $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} g$ and we want to fit a density from the parametric model $\mathcal{F} := \{f_{\theta} : \theta \in \Theta\}$.
- We can minimize $\rho_{\alpha}(g, f_{\theta})$ over θ in order to get a “reasonable” estimate.

Introduction (Contd.)

- In order to get the minimizer, we need to estimate $\rho_\alpha(g, f_\theta)$. Note that we do not need the first term and the third term need not be estimated. For the second term, realizing that it is an expectation we can estimate it by

$$\frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i).$$

- Minimum DPD estimator corresponding to the parameter $\alpha(> 0)$ is defined by

$$\hat{\theta}_\alpha := \operatorname{argmin}_\theta \int f_\theta^{1+\alpha}(x) dx - \frac{1+\alpha}{\alpha} \frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i).$$

- For $\alpha = 0$, it was defined by taking limit as $\alpha \rightarrow 0$ of $\rho_\alpha(g, f_\theta)$. $\hat{\theta}_0$ corresponds to the maximum likelihood estimator.

Properties

- The main advantage of DPD estimator over general robust M-estimator is that the properties of the estimator like robustness and efficiency can be tuned by changing the value of (scalar) tuning parameter α .
- As α increases from 0, the estimator becomes more robust and less efficient.
- The case $\alpha = 1$ corresponds to L_2 estimator which has relative efficiency of about 50% in normal model. Hence, usually we consider $0 < \alpha < 1$.
- DPD estimator has been extensively studied in case of estimation in regular parametric models including the case of censored data and non-homogeneous observations (regression). See, for example, Ghosh and Basu (2014).

DPD Estimation

- Considering the model $y_i = m_0(x_i) + \epsilon_i$ with $\epsilon_1, \epsilon_2, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, 1)$, the corresponding DPD estimators of m_0 would be given as follows.
- **DPD-P1**: minimize over $f \in C^k$,

$$-\frac{1}{n\alpha} \sum_{i=1}^n \exp\left(-\frac{\alpha}{2}[y_i - f(x_i)]^2\right) + \frac{1}{\alpha} + \lambda_n \int_{[0,1]^d} f^{(k)}(x)^2 dx.$$

- **DPD-P2**: minimize over $\theta \in \mathbb{R}$ for each $x \in [0, 1]^d$,

$$\frac{1}{n} \sum_{i=1}^n W_i(x) \left[\frac{1}{\alpha} - \frac{1}{\alpha} \exp\left(-\frac{\alpha}{2}[y_i - \theta]^2\right) \right],$$

to get $\hat{m}_\alpha(x)$.

Remarks

- Note that if $\alpha \rightarrow 0$, then both the objective functions coincide with those of **P1** and **P2**.
- Objective functions in **DPD-P1** and **DPD-P2** can be accordingly modified with respect to the error distribution.
- That is, we can write

$$\int f_m^{\alpha+1}(y|x_i) dy - \frac{1+\alpha}{\alpha} f_m^\alpha(y_i|x_i),$$

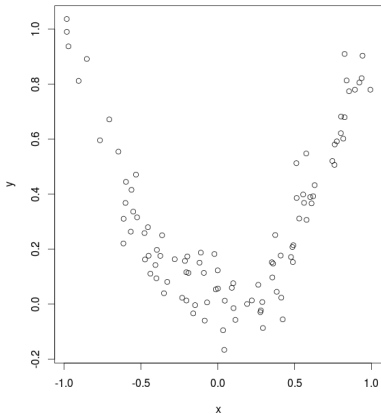
in place of $[y_i - m(x_i)]^2$ in **P1** and **P2**. This would allow us to study the efficiency properties of the non-parametric estimator \hat{m}_α and tune α accordingly.

Properties of \hat{m}_α from DPD-P1

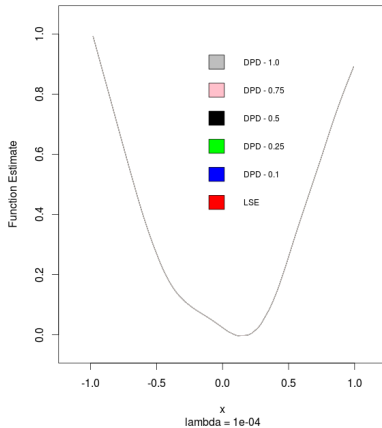
- The minimizer of the objective function in DPD-P1 is a natural spline of degree $2k - 1$. Hence an iterative algorithm similar to that given in Reinsch (1967) can be used to calculate the estimator.
- This algorithm is observed to be numerically more stable than the iteratively reweighed least squares algorithm.
- The general results of Cox (1983) proves that the estimator \hat{m}_α leads to optimal rate of convergence of $\|\hat{m}_\alpha - m_0\|_2$ for every α . Also see Oh, Nychka and Lee (2007) for more details.
- Simulations under different contaminations are as follows.

Contamination: 0% and Sample Size: 100

Error Contamination: 0%

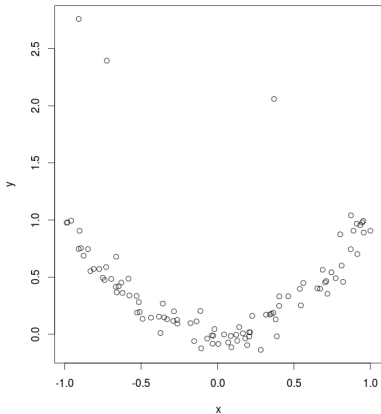


Function Estimate by DPD and LS

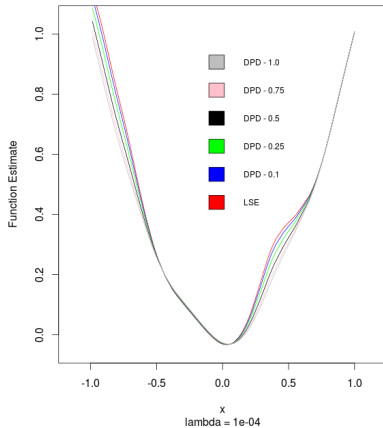


Contamination: 1% and Sample Size: 100

Error Contamination: 1%

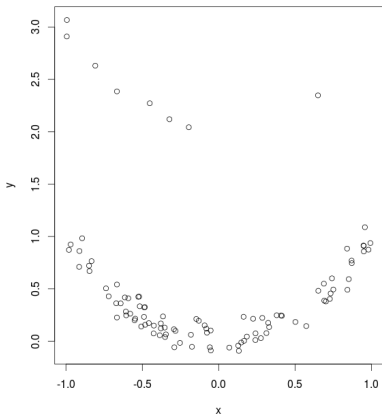


Function Estimate by DPD and LS

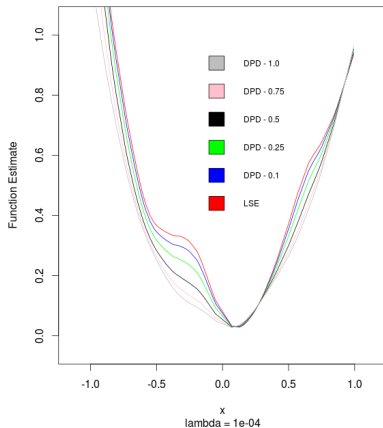


Contamination: 5% and Sample Size: 100

Error Contamination: 5%

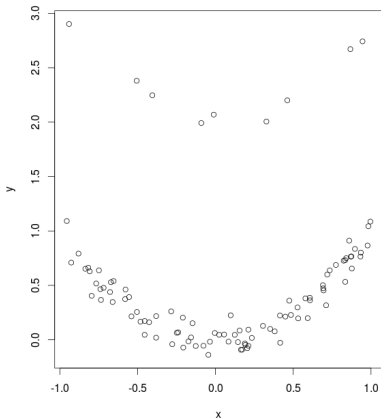


Function Estimate by DPD and LS

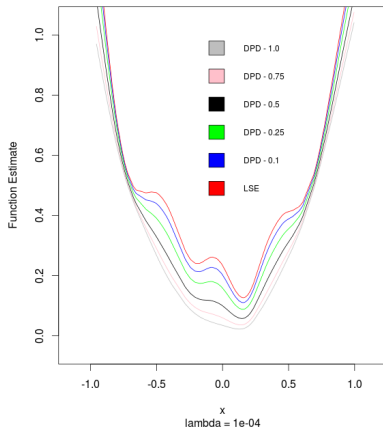


Contamination: 10% and Sample Size: 100

Error Contamination: 10%

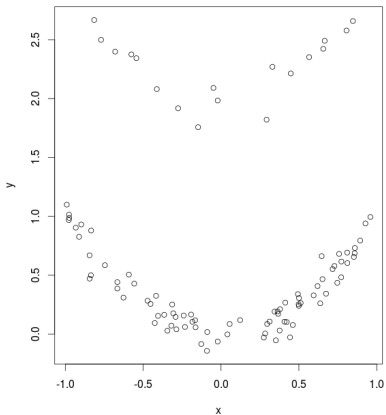


Function Estimate by DPD and LS

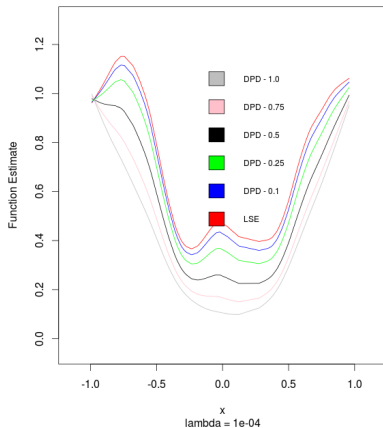


Contamination: 25% and Sample Size: 100

Error Contamination: 25%



Function Estimate by DPD and LS



Properties of \hat{m}_α from **DPD-P2**

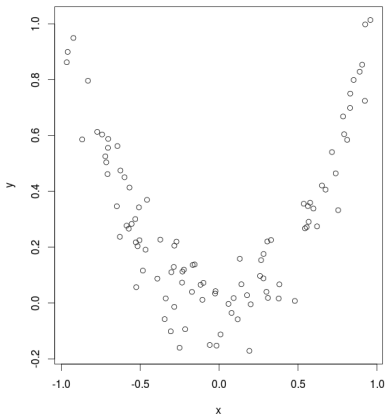
- The objective function here naturally leads to an iterative algorithm. For $\alpha = 0$, the estimator coincides with the non-parametric regression function estimator based on the weights $W_i(x)$, i.e.,

$$\sum_{i=1}^n W_i(x) Y_i / \sum_{i=1}^n W_i(x).$$

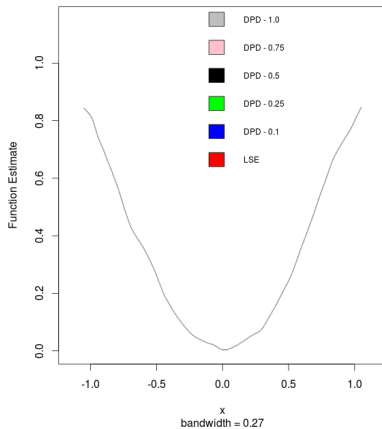
- Pointwise properties of these estimators were extensively studied in literature. Their general results prove that under certain regularity conditions, related to the weight function, there exists c_n (depending on weight function) such that for any set of points t_1, t_2, \dots, t_k , $c_n(\hat{m}_\alpha(t_{1:k}) - m_0(t_{1:k})) \xrightarrow{\mathcal{L}} N(0, V_\alpha(t_{1:k}))$. See Boente and Fraiman (1989) for more details.

Contamination: 0% and Sample Size: 100

Error Contamination: 0%

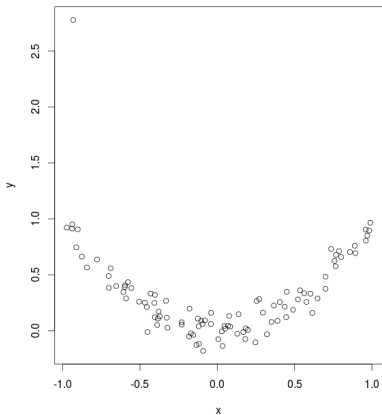


Function Estimate by DPD and LS

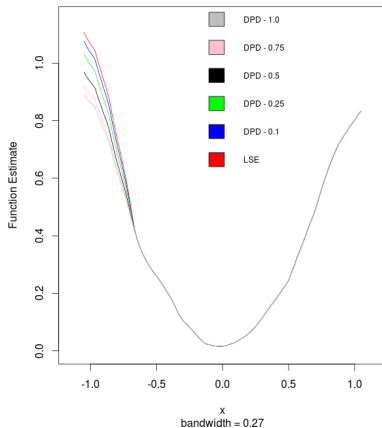


Contamination: 1% and Sample Size: 100

Error Contamination: 1 %

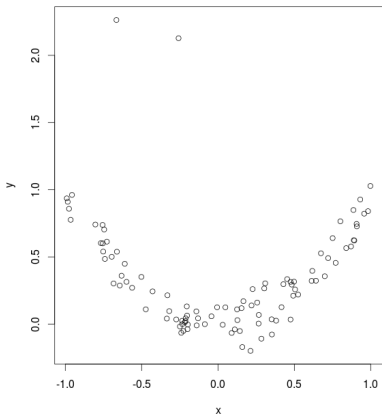


Function Estimate by DPD and LS

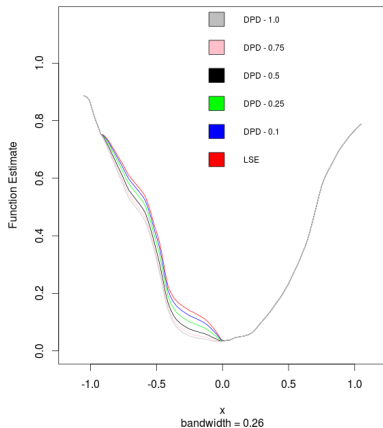


Contamination: 5% and Sample Size: 100

Error Contamination: 5 %

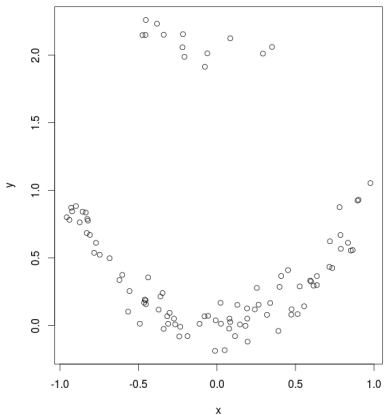


Function Estimate by DPD and LS

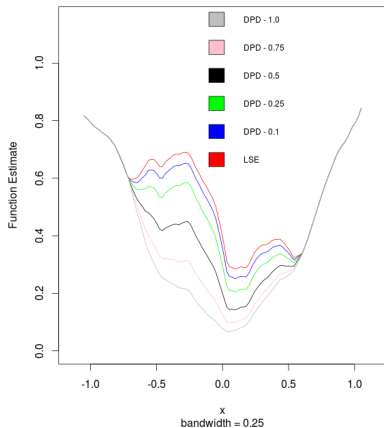


Contamination: 10% and Sample Size: 100

Error Contamination: 10 %

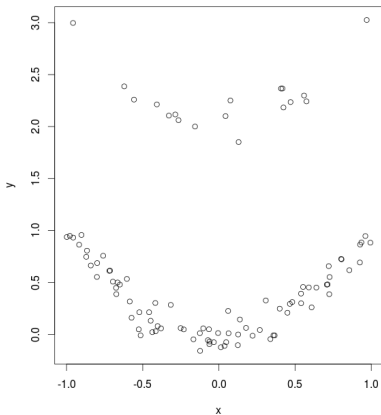


Function Estimate by DPD and LS

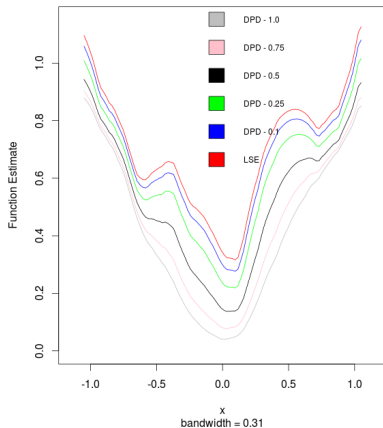


Contamination: 25% and Sample Size: 100

Error Contamination: 25 %








Function Estimate by DPD and LS







Conclusions

- Similar results can be obtained by including variance function (both homoscedastic and heteroscedastic). In case of variance function estimation, using the joint normality result, we can get confidence bands for \hat{m}_α .
- Choice of λ_n are critical in getting the asymptotic results. Choice of α is critical in getting a good estimate (robust or efficient depending on the data).
- Choice of λ_n can be done by cross-validation. But asymptotic analysis related to data-driven choices of α and λ_n are yet to be done.
- An R-Package for these two methods is currently under preparation.

References

-  Cleveland, William S. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74 (1979), no. 368, 829–836.
-  Utreras, Florencio I. On computing robust splines and applications. *SIAM J. Sci. Statist. Comput.* 2 (1981), no. 2, 153–163.
-  Härdle, Wolfgang Robust regression function estimation. *J. Multivariate Anal.* 14 (1984), no. 2, 169–180.
-  Härdle, W.; Tsybakov, A. B. Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.*, 16 (1988) no. 1, 120–135.
-  Oh, Hee-Seok; Nychka, Douglas W.; Lee, Thomas C. M. The role of pseudo data for robust smoothing with application to wavelet regression. *Biometrika*, 94 (2007), no. 4, 893–904.

References (Contd.)

-  Chaudhuri, Probal; Dewanji, Anup On a likelihood-based approach in nonparametric smoothing and cross-validation. *Statist. Probab. Lett.* 22 (1995), no. 1, 7–15.
-  Boente, Graciela; Fraiman, Ricardo Robust nonparametric regression estimation. *J. Multivariate Anal.* 29 (1989), no. 2, 180–198.
-  Reinsch, Christian H. Smoothing by spline functions. I, II. *Numer. Math.* 10 (1967), 177–183; *ibid.* 16 (1970/71), 451–454.
-  Basu, Ayanendranath; Harris, Ian R.; Hjort, Nils L.; Jones, M. C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85 (1998), no. 3, 549–559.

Density Estimation based on DPD

- Density estimation using DPD can be done in three ways.
- one of these is similar to the penalized likelihood estimation of density. This proposal requires minimization of

$$\int f^{1+\alpha}(x)dx - \frac{1}{n} \sum_{i=1}^n f^{\alpha}(X_i) + \lambda \int f''(x)^2 dx.$$

over $f \in C^2$.

- Other proposals use the regression function estimate derived using **DPD-P1** and **DPD-P2**. These use the **root-unroot** method proposed by Brown et al. (2010) showing equivalence of density estimation problem and regression function estimation problem.
- This method can be described as follows.

Root – Unroot method

- *Binning*: Divide $\{X_i\}$ into T equal length intervals between 0 and 1. Let Q_1, Q_2, \dots, Q_T be the number of observations in each of the intervals.
- *Root Transform*: Let $Y_i = \sqrt{\frac{T}{n}} \sqrt{Q_i + \frac{1}{4}}, i = 1, \dots, T$, and treat $Y = (Y_1, Y_2, \dots, Y_T)$ as the new equispaced sample for a non-parametric regression problem.
- *Regression Set up*: Let $h(x) = \sqrt{f(x)}$ and q_j be the mid-point of j th interval. Then

$$Y_j \approx h(q_j) + \sigma \epsilon_j,$$

where $\epsilon_j \sim N(0, 1)$ and $\sigma = \sqrt{\frac{T}{4n}}$.

Root – Unroot method

- *Non-parametric Regression*: Apply any non-parametric regression procedure to (q_j, Y_j) to obtain an estimate \hat{h} of \sqrt{f} .
- *Unroot*: The density function f is estimated by $\hat{f} = \hat{h}^2$.
- *Normalization*: The estimator \hat{f} given in Step 4 may not integrate to 1. Set

$$\tilde{f} = \frac{\hat{f}}{\int_0^1 \hat{f}(t) dt},$$

and use \tilde{f} as the final estimator.

Thank You!