GR^{2}T Vs L$_2$E with nuisance scale

Rozenn Dahyot
School of Computer Science and Statistics
Trinity College Dublin, Ireland
Email: Rozenn.Dahyot@tcd.ie

***Abstract*—We compare the objective functions used by GR^{2}T [1] and the L$_2$E estimator [2] that have both been proposed for robust parameter estimation. We show their similarity when estimating location parameters. Of particular interest is their ability for dealing with the scale parameter that is often unknown and acts as a nuisance parameter. Both techniques are tested experimentally for regression (e.g. to find patterns such as line and circle in noisy datasets) and for registration between datasets.**

I. Introduction

One major difficulty in robust inference is to deal with the unknown standard deviation of the inliers also known as the scale. Assuming that the correct parameter of interest (location parameter) has been recovered, a robust estimate of the scale can be found based on the distribution of the residuals [3]. However, a poor estimate of the location parameter leads to a poor scale parameter estimate and vice versa. Two frameworks, GR^{2}T [1] and L$_2$E [2], have recently been introduced for dealing jointly with both scale and location parameters. We first point out their similarities and differences in section II. We extend GR^{2}T formulation to the problem of registration (paragraph III) and we propose to take advantage of the Bayesian nature of GR^{2}T to add a prior distribution for the scale (section IV). Section V compares experimentally both GR^{2}T and L$_2$E for both regression and registration. Section VI concludes with potential improvements on the modelling the scale prior.

II. L$_2$E and GR^{2}T

A. Euclidian distance L_2/L_2E between pdfs

1) L_2 for Registration: Considering two datasets $\{x^{(i)}\}_{i=1,\cdots,N_x}$ and $\{y^{(j)}\}_{j=1,\cdots,N_y}$, Jian et al. [4] proposes to find the transformation t that registers the first dataset onto the second, such that the Euclidian distance between the kernel density estimate noted $\hat{p}_{t(x)}$ computed using $\{t(x^{(i)})\}_{i=1,\cdots,N_x}$ and a kernel density estimate noted $\hat{p}_y$ computed using observations $\{y^{(j)}\}_{j=1,\cdots,N_y}$ is minimised. This Euclidian distance corresponds to:

$$\begin{aligned} L_2(t) &= \|\hat{p}_y - \hat{p}_{t(x)}\|^2 = \int (\hat{p}_y(y) - \hat{p}_{t(x)}(y))^2 \, dy \\ &= \|\hat{p}_y\|^2 + \|\hat{p}_{t(x)}\|^2 - 2\,\langle \hat{p}_y | \hat{p}_{t(x)} \rangle \end{aligned} \tag{1}$$

and the transformation t is estimated with:

$$\begin{aligned} \hat{t} &= \arg\min_t L_2(t) \\ &= \arg\min_t \left\{ L_2E(t) = \|\hat{p}_{t(x)}\|^2 - 2\,\langle \hat{p}_y | \hat{p}_{t(x)} \rangle \right\} \end{aligned} \tag{2}$$

since the term $\|\hat{p}_y\|$ does not depend on t. When the transformation t to estimate is a rigid transformation then the term $\|\hat{p}_{t(x)}\|$ also does not depend on t [4], in which case it is equivalently found by maximising the kernel correlation [5]:

$$\hat{t} = \arg\max_t \langle \hat{p}_y | \hat{p}_{t(x)} \rangle \tag{3}$$

Note that when using Gaussian Kernel density estimates for $\hat{p}_y$ and $\hat{p}_{t(x)}$, integrals in eq. (1) or (3) are solved explicitly [4]. While L$_2$E or kernel correlation are not objective functions originally able to include prior information about t, several recent works have proposed to include an additive regularisation term to L$_2$/L$_2$E to constraint its estimation [6], [7].

2) L_2E objective function for Regression: Consider the following equation:

$$F(x, \theta) = \epsilon \sim p_{\epsilon|\nu}(\epsilon) \tag{4}$$

where θ is the latent random variable of interest that we wish to infer, F is a link function (linear or not) relating the observed variable x with θ, and the noise ϵ has a chosen distribution $p_{\epsilon|\nu}$ that is centered on zero and depends on a (scale) nuisance parameter ν. Scott [2] proposed to estimate θ and the scale ν by minimising L$_2$E between the chosen model for the errors $p_{\epsilon|\nu}$ and the empirical pdf $\hat{p}_\epsilon$ defined as:

$$\hat{p}_\epsilon(\epsilon) = \frac{1}{N} \sum_{i=1}^{N} \delta(\epsilon - \epsilon^{(i)}) \tag{5}$$

where $\delta(\cdot)$ is the Dirac kernel, and $\epsilon^{(i)} = F(x^{(i)}, \theta)$ is the residual computed at θ using the observations $\{x^{(i)}\}_{i=1,\cdots,N}$ collected for the variable x. The location θ and scale ν are then estimated as follow:

$$(\hat{\theta}, \hat{\nu}) = \arg\min_{\theta,\nu} \left\{ L_2E(\theta, \nu) = \|p_{\epsilon|\nu}\|^2 - 2\,\langle p_{\epsilon|\nu} | \hat{p}_\epsilon \rangle \right\} \tag{6}$$

The dependence over θ only appears thanks to the empirical density $\hat{p}_\epsilon$ (through the residuals), hence when ν is fixed, θ is estimated by:

$$\begin{aligned} \hat{\theta} &= \arg\min_\theta L_2E(\theta, \nu) \\ &= \arg\max_\theta \left\{ \langle p_{\epsilon|\nu} | \hat{p}_\epsilon \rangle \right\} \\ &= \arg\max_\theta \left\{ \tfrac{1}{N} \textstyle\sum_{i=1}^{N} p_{\epsilon|\nu}(\epsilon^{(i)}) \right\} \end{aligned} \tag{7}$$

The estimation for both θ and ν with L$_2$E can be rewritten:

$$(\hat{\theta}, \hat{\nu}) = \arg\max_{\theta,\nu} \left\{ \langle p_{\epsilon|\nu} | \hat{p}_\epsilon \rangle - \frac{1}{2} \|p_{\epsilon|\nu}\|^2 \right\} \tag{8}$$

The term $\frac{1}{2}\|p_{\epsilon|\nu}\|^2$ can be thought as a barrier function [8] to prevent the scale estimate $\hat{\nu}$ to be zero. Moreover the estimate $\hat{\nu}$ is exactly the true parameter ν_T when the empirical estimate $\hat{p}_\epsilon$ (eq. 5) converges exactly towards the model $p_{\epsilon|\nu_T}$. Note that when outliers occur, $\hat{p}_\epsilon$ will not converge towards a good model from the model family $p_{\epsilon|\nu}$. L$_2$E robustness in finding θ is then observed for a well chosen fixed scale [2].

B. Generalised Relaxed Radon Transform (GR^{2}T)

The Generalised Relaxed Radon Transform (GR^{2}T) has recently been proposed for robust regression [1] and it is augmenting the original problem (4) by adding an auxiliary variable λ as follow:

$$\begin{cases} \lambda + F(x,\theta) = \epsilon \sim p_{\epsilon|\nu}(\epsilon) \\ \lambda = 0 \end{cases} \tag{9}$$

The problem stated in equation (9) is equivalent to the original equation (4) : GR^{2}T proposes to use the first equation in (9) to compute estimates of the pdfs $p_{\lambda|\theta}$ and $p_{\lambda\theta}$ while the second equation is used for narrowing down the search in the latent space to the special case of interest when $\lambda = 0$ [1]. The joint density function $p_{\lambda\theta}$ corresponds to:

$$\begin{aligned} p_{\theta\lambda}(\theta,\lambda) &= \int p_{\lambda\theta|x}(\lambda,\theta|x)\ p_x(x)\ dx \\ &= \langle p_{\lambda\theta|x}|\ p_x\rangle \\ &= \langle p_{\lambda|x\theta}\ p_{\theta|x}|\ p_x\rangle \end{aligned} \tag{10}$$

Given equation (9), the conditional $p_{\lambda|x\theta}$ is defined with the noise model $p_{\epsilon|\nu}$ as follow:

$$p_{\lambda|x\theta}(\lambda|x,\theta) = p_{\epsilon|\nu}(\lambda + F(x,\theta)) \tag{11}$$

Note that when the error distribution $p_{\epsilon|\nu}$ is the Dirac density function $\delta(\epsilon)$ (with $\nu = 0$), the probability density function $p_{\Theta\lambda}$ corresponds to the Generalised Radon Transform [9]. If no prior is available to model the conditional $p_{\theta|x}$, one can assume independence $p_{\theta|x} = p_\theta$ leading to:

$$p_{\lambda\theta}(\lambda,\theta) = p_\theta(\theta)\ \underbrace{\langle p_{\lambda|x\theta}|\ p_x\rangle}_{p_{\lambda|\theta}(\lambda|\theta)} \tag{12}$$

Using the observations $\{x^{(i)}\}_{i=1,\cdots,N}$, the empirical probability density function of x can be computed:

$$\hat{p}_x = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x^{(i)})$$

and the conditional $p_{\lambda|\theta}$ can then be estimated by the empirical average [1], [10]:

$$\hat{p}_{\lambda|\theta}(\lambda|\theta) = \langle p_{\lambda|\theta x}|\ \hat{p}_x\rangle = \frac{1}{N}\sum_{i=1}^{N} p_{\epsilon|\nu}(\epsilon^{(i)}) \tag{13}$$

Inference about θ in the case of interest $\lambda = 0$ can then be done using the estimated posterior:

$$\hat{p}_{\theta|\lambda}(\theta|\lambda=0) = \frac{p_\theta(\theta)\ \hat{p}_{\lambda|\theta}(\lambda=0|\theta)}{\int p_\theta(\theta)\ \hat{p}_{\lambda|\theta}(\lambda=0|\theta)\ d\theta} \tag{14}$$

A maximum a posteriori estimate of the location parameter can be computed as:

$$\hat{\theta} = \arg\max_\theta \{\hat{p}_{\theta|\lambda}(\theta|\lambda=0) \propto p_\theta(\theta)\ \hat{p}_{\lambda|\theta}(\lambda=0|\theta)\} \tag{15}$$

or simply using the estimated conditional $\hat{p}_{\lambda|\theta}$ when no prior p_θ is available e.g.:

$$\hat{\theta} = \arg\max_\theta \{\hat{p}_{\lambda|\theta}(\lambda=0|\theta)\} \tag{16}$$

Augmenting equation (9) with an additive auxiliary variable λ allows to not care about the potentially complex nature of the function F (e.g. linear or non linear) in the modelling and also allow the usage of a prior on the parameter of interest θ. Note how the estimate of the conditional $p_{\lambda|\theta}$ in equation (13) is identical to L$_2$E in equation (7) when the scale is fixed. Note also that GR^{2}T is a Bayesian modelling that allows the inclusion of prior distribution about the latent variable θ and contrary to recent works adding regularisation terms to L$_2$ [6], [7], the prior in GR^{2}T is instead multiplied to $p_{\lambda|\theta}$. Reference [1] explains how GR^{2}T encapsulates the following robust frameworks: the Hough transform [11]–[14], M-estimators [15] and Generalized Projection Based M-Estimators [16], [17].

III. GR^{2}T FOR REGISTRATION

In this section, we reformulate briefly GR^{2}T for registration. Consider two datasets $\{x^{(i)}\}_{i=1,\cdots,N_x}$ and $\{y^{(j)}\}_{j=1,\cdots,N_y}$ observations for the two random variables x and y respectively. Assume the following relationship beween x and y parameterised by a latent variable θ:

$$\lambda + F(x,y,\theta) = \epsilon \sim p_\epsilon(\epsilon) \tag{17}$$

where λ is an auxiliary variable and the case of interest is when $\lambda = 0$. A standard model for registration is to express the transformation t for mapping x onto y in parametric form:

$$F(x,y,\theta) = y - t(x,\theta) \tag{18}$$

For instance, we will consider a translation in the experiment in section V-B):

$$F(x,y,\theta) = y - (x + \theta) \tag{19}$$

Using the same formulation for GR^{2}T as presented in section II-B, the conditional $p_{\lambda|\theta}$ corresponds to:

$$p_{\lambda|\theta}(\lambda|\theta) = \langle p_{\lambda|\theta xy}|\ p_{xy}\rangle$$

The question is how to compute an empirical estimate p_{xy}. If the observations $\{x^{(i)}\}_{i=1,\cdots,N_x}$ and $\{y^{(j)}\}_{j=1,\cdots,N_y}$ can be grouped into a set of correspondences $\{(x^{(k)},y^{(k)})\}_{k=1,\cdots,K}$, then the following estimate can be used.

$$\hat{p}_{xy}(x,y) = \frac{1}{K}\sum_{k=1}^{K}\delta(x - x^{(k)})\ \delta(y - y^{(k)})$$

However, when no correspondences are available, one can assume independence between x and y and use the following empirical pdf:

$$\begin{aligned} &\hat{p}_{xy}(x,y) \\ &= \left(\tfrac{1}{N_x}\textstyle\sum_{i=1}^{N_x}\delta(x - x^{(i)})\right)\ \left(\tfrac{1}{N_y}\textstyle\sum_{j=1}^{N_y}\delta(y - y^{(j)})\right) \\ &= \tfrac{1}{N_x\ N_y}\textstyle\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\delta(x - x^{(i)})\ \delta(y - y^{(j)}) \end{aligned} \tag{20}$$

leading to

$$\hat{p}_{\lambda|\theta}(\lambda|\theta) = \frac{1}{N_x\ N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} p_{\epsilon|\nu}(\epsilon^{(i,j)}) \tag{21}$$

with the residuals $\epsilon^{(i,j)} = \lambda + F(x^{(i)},y^{(j)},\theta),\ \forall i,j$. Experimental results shown in V-B have been computed without correspondence between the observations using the expression (21) augmented with a prior distribution about the scale ν explained next.

IV. GR^{2}T WITH SCALE PRIOR

When considering ν as a random variable equation (11) is more accurately rewritten:

$$p_{\lambda|\theta x\nu}(\lambda|\theta, x, \nu) = p_{\epsilon|\nu}(\epsilon|\nu) \tag{22}$$

Assuming the scale ν independent of θ and x, the location and scale can be estimated by:

$$(\hat{\theta}, \hat{\nu}) = \arg\max_{\theta,\nu} \left\{ \hat{p}_{\lambda\nu|\theta}(\lambda, \nu|\theta) = p_\nu(\nu)\ \hat{p}_{\lambda|\nu\theta}(\lambda|\nu, \theta) \right\} \tag{23}$$

The log-normal distribution with zero mean is chosen as the prior distribution for the scale ν:

$$p_\nu(\nu) = \frac{1}{\nu\sqrt{2\pi}\gamma} \exp\left(-\frac{(\log\nu)^2}{2\gamma^2}\right), \quad \nu > 0,\ \gamma > 0 \tag{24}$$

γ is an hyperparameter controling the shape of the prior: large γ (e.g. $\gamma = 4$ is used in all experiments) favorises small scales ν (more suitable for inliers) yet preventing ν to be zero. Optimisation is performed by augmenting γ progressively to 4 to avoid local solutions.

Note that when the scale ν is very large ($\nu \to \infty$) then the pdf of the error can be approximated using Taylor expansion (with $\frac{\epsilon}{\nu}$ near zero) as follow:

$$p_{\epsilon|\nu}(\epsilon|\nu) = \frac{1}{\sqrt{2\pi}\nu} \exp\left(-\frac{\epsilon^2}{2\nu^2}\right) \simeq \frac{1}{\sqrt{2\pi}\nu}\left(1 - \frac{\epsilon^2}{2\nu^2}\right)$$

Hence estimating θ by maximising $\hat{p}_{\lambda|\theta\nu}$ when $\nu \to \infty$ is the same as minimising the sum of square errors and therefore leads to the standard maximum likelihood (ML) solution. Estimation using GR^{2}T and L$_2$E (eq. 8) are performed using a gradient ascent algorithm using ML estimate $(\hat{\theta}_{ML}, \hat{\nu}_{ML})$ as the initial guess.

V. EXPERIMENTAL RESULTS

We compare experimentally three approaches to estimate the parameter θ of interest. The first is the standard Maximum Likelihood (ML) approach that is known to not be robust to outliers. This solution serves as an initial guess for both $\mathcal{L}_2E$ and GR^{2}T. The second is $\mathcal{L}_2E$ that is known to be more robust to outliers in particular when the scale is well set: in this paper the scale is also estimated by minimising the $\mathcal{L}_2E$ cost function. The last method corresponds to GR^{2}T with a scale prior.

A. Regression

To assess our algorithm we used several datasets published by Toldo et al. [18] that include many outliers and pseudo-outliers. The equations used to find the main pattern are:

$$\lambda + (x_1 - (\theta_1 + \theta_2\ x_2)) = \epsilon \quad \text{(line)}$$

and

$$\lambda + \sqrt{(x_1 - \theta_1).^2 + (x_2 - \theta_2).^2} - \theta_3 = \epsilon \quad \text{(circle)}$$

Figure 1 shows several results: while L$_2$E and ML systematically fail, GR^{2}T manages to lock on a line in (b) and (c) but is getting trapped in a solution with a higher scale in the dataset (d). GR^{2}T also manages to find a circle in (a) while ML and L$_2$E fails.

B. Registration

We compare the three estimation techniques to recover the translation parameters to register two point sets. The first point cloud $\{x^{(i)}\}_{i=1,\cdots,N_x}$ corresponds to the 2D fish data (composed of $N_x = 98$ 2D points) [4]. The second dataset $\{y^{(j)}\}_{j=1,\cdots,N_y}$ is a translated version of the first point set (ground truth translation parameters equal to $(-1, -1)$) with some noise. The equation used for registration here is:

$$\lambda + \|y - (x + \theta)\| = \epsilon \tag{25}$$

with $y \in \mathbb{R}^2$, $x \in \mathbb{R}^2$, the translation parameter $\theta \in \mathbb{R}^2$ to estimate, and the noise $\epsilon \in \mathbb{R}$ follows a Normal distribution with mean zero and variance ν^2 (the prior for ν is again the log-normal distribution). Figure 2 compares the methods L$_2$E and GR^{2}T for aligning these two point sets for various levels of contamination (outliers), noise on the inliers, and missing data. The results obtained by ML is not shown at it fails systematically when outliers occur. Adding the prior for the scale in GR^{2}T allows to control that ν is not over-estimated as this would lead to a bad estimate for θ. As a consequence GR^{2}T performs better than L$_2$E.

VI. DISCUSSION AND FUTURE WORK

This paper has presented how GR^{2}T can be efficiently used to register datasets. Secondly prior information about the scale can also be added in this Bayesian framework to help the robust estimation of the location parameter. While the rigid transformation for registration of 2D point sets has been modelled using equation (25) in this paper, the following alternative system of equations would have also been suitable:

$$(\lambda + F(x, y, \theta) = \epsilon) \equiv \left(\begin{cases} \lambda_1 + y_1 - x_1 - \theta_1 = \epsilon_1 \\ \lambda_2 + y_2 - x_2 - \theta_2 = \epsilon_2 \end{cases} \right)$$

with notations $y = (y_1, y_2)$ and $x = (x_1, x_2)$, $\theta = (\theta_1, \theta_2)$ and in this case $\epsilon = (\epsilon_1, \epsilon_2)$ is a random vector in $\mathbb{R}^2$. $p_{\epsilon|\nu}$ can then be modelled with a bivariate Normal distribution controlled by a covariance matrix ν. Future work will look at choosing more suitable prior distributions for these nuisance parameters ν, and also to model for instance heteroscedastic noise [19].

ACKNOWLEDGMENTS

This research has been partly supported by EU FP7-PEOPLE-2013-IAPP grant GRAISearch (612334).

REFERENCES

[1] R. Dahyot and J. Ruttle, "Generalised relaxed radon transform (gr2t) for robust inference," *Pattern Recognition*, vol. 46, no. 3, 2013.

[2] D. W. Scott, "Parametric statistical modeling by minimum integrated square error," *Technometrics*, vol. 43, no. 3, pp. pp. 274–285, 2001. [Online]. Available: http://www.jstor.org/stable/1271214

[3] H. Wang and D. Suter, "Robust adaptive-scale parametric model estimation for computer vision," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, November 2004.

[4] B. Jian and B. Vemuri, "Robust point set registration using gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1633 – 1645, 2011.

[5] Y. Tsin and T. Kanade, "A correlation-based approach to robust point set registration," in *European Conference in Computer Vision ECCV*, 2004, pp. 558–569.

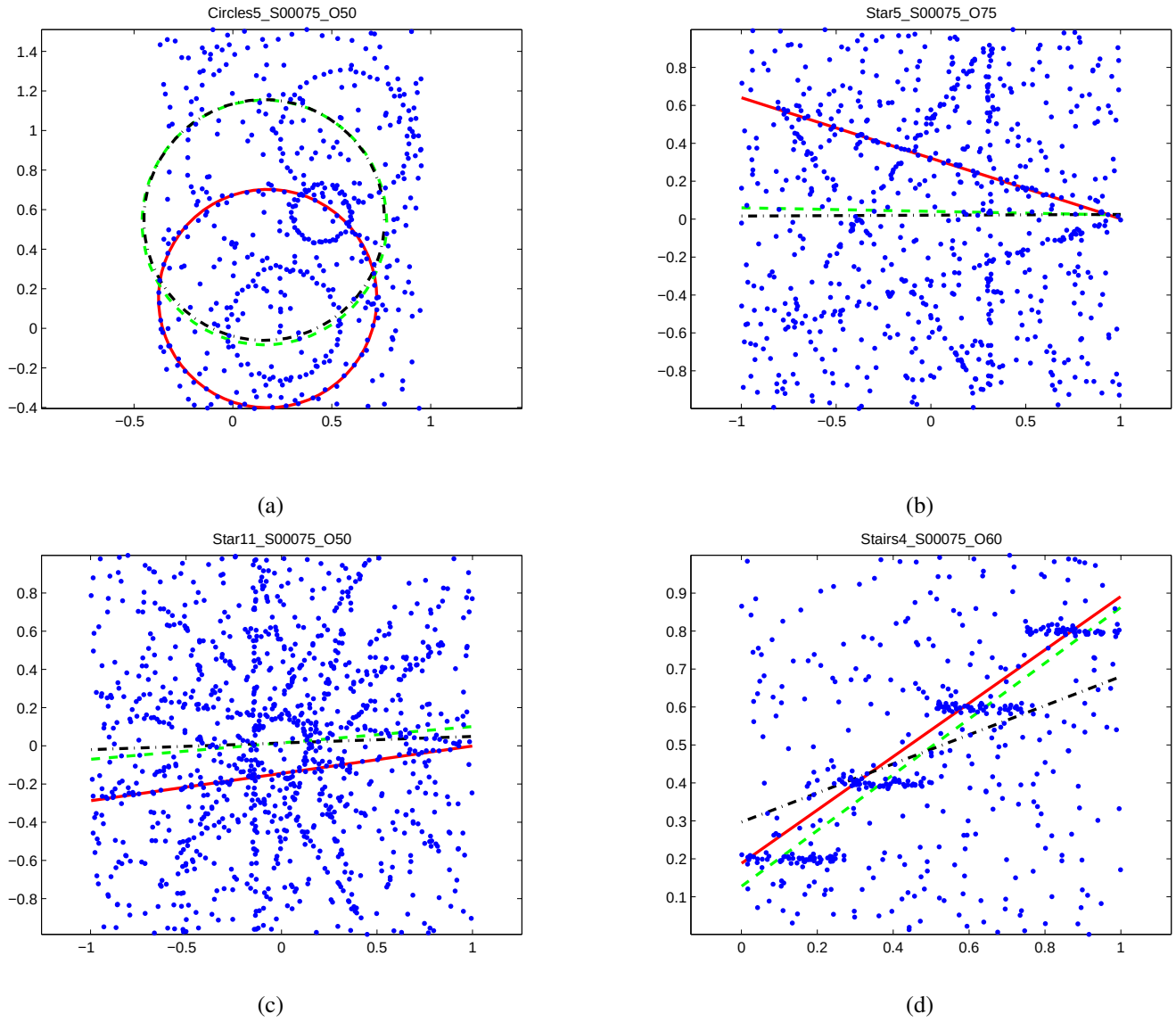

Fig. 1. **Robust Regression with Outliers and Pseudo-Outliers**: GR^{2}T with Scale Prior (red), L$_2$E (green, dashed) , ML (black, dot-dashed) (using Jlinkage datasets [18]).

[6] C. Arellano and R. Dahyot, "Shape model fitting algorithm without point correspondence," in *20th European Signal Processing Conference (Eusipco)*, Bucharest, Romania, August, 27-31 2012, pp. 934–938.

[7] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. L. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA USA, June 23-28 2013.

[8] M. H. WRIGHT, "The interior-point revolution in optimization: History, recent developments, and lasting consequences," *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*, vol. 42, no. 1, pp. 39–56, 2004.

[9] K. Denecker, J. V. Overloop, and F. Sommen, "The general quadratic radon transform," *Inverse problems*, no. 14, pp. 615–633, 1998.

[10] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer Verlag, 1999.

[11] P. Hough, "Methods of means for recognising complex patterns," US Patent 3069654, 1962.

[12] N. Kiryati, Y. Eldar, and A. M. Bruckstein, "A probabilistic hough transform," *Pattern Recognition*, vol. 24, no. 4, pp. 303–316, 1991.

[13] J. Princen, J. Illingworth, and J. Kittler, "A formal definition of the hough tranform: Properties and relationships," *Journal of Mathematical Imaging and Vision*, vol. 1, no. 2, pp. 153–168, 1992.

[14] C. F. Olson, "Constrained hough transform for curve detection," *Computer Vision and Image Understanding*, vol. 73, no. 3, March 1999.

[15] P. Huber, *Robust Statistics*. John Wiley and Sons, 1981.

[16] S. Mittal, S. Anand, and P. Meer, "Generalized projection based m-estimator: Theory and applications," in *Computer Vision and Pattern Recognition Conference*, Colorado Springs, CO, June 2011, pp. 2689–2696.

[17] ——, "Generalized projection based m-estimator." *IEEE transactions on Pattern analysis and machine intelligence*, 2012.

[18] R. Toldo and A. Fusiello, "Robust multiple structures estimation with j-linkage," in *European Conference on Computer Vision (ECCV) 2008*, Marseille, France, 2008. [Online]. Available: http://profs.sci.univr.it/~fusiello/demo/jlk/

[19] C. M. Crainiceanu, D. Ruppert, R. J. Carroll, A. Joshi, and B. Goodner, "Spatially adaptive bayesian penalized splines with heteroscedastic errors," *Journal of Computational and Graphical Statistics*, vol. 16, no. 2, pp. 265–288, 2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/106186007X208768

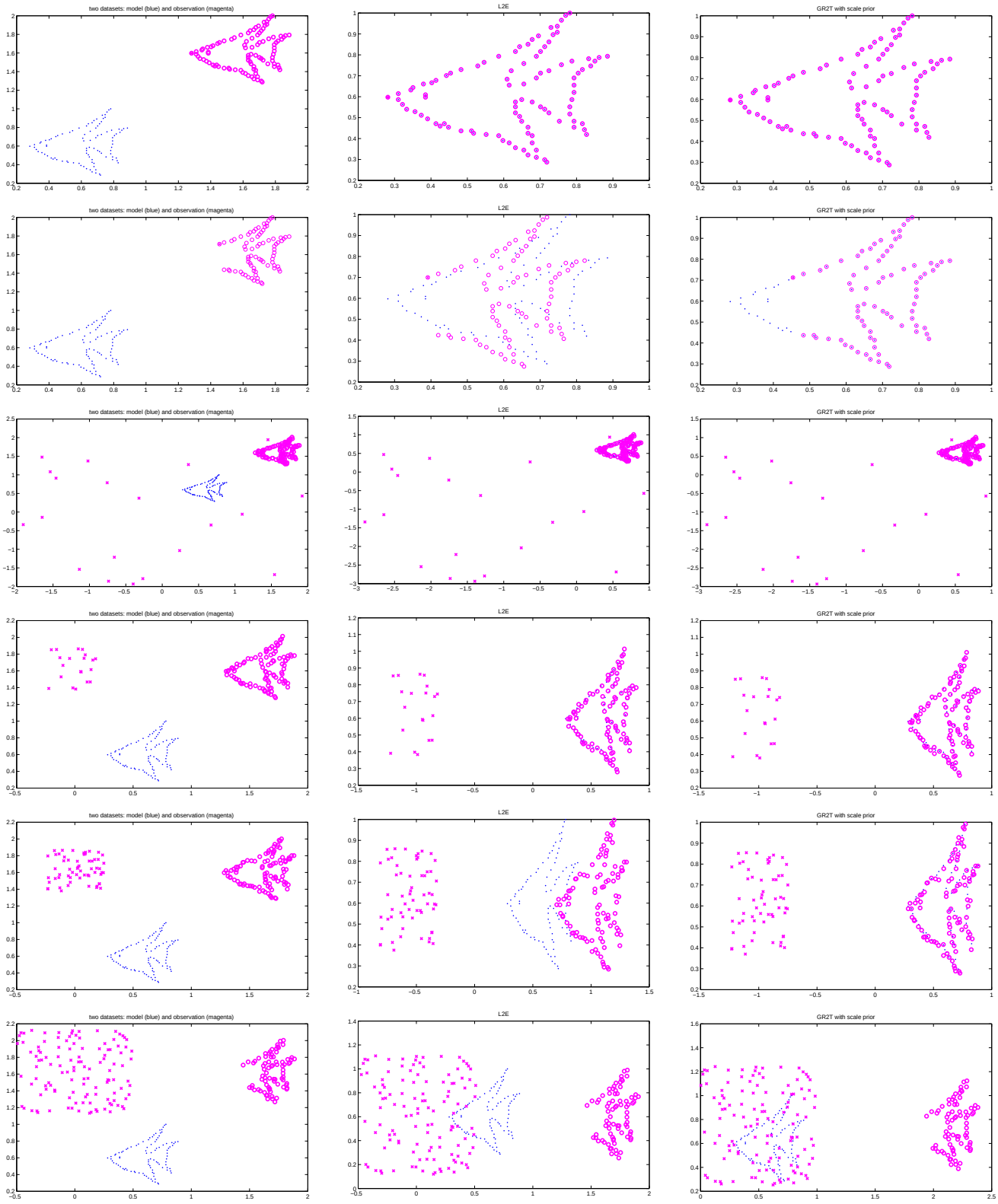

Fig. 2. Rigid Registration (translation) of 2D point clouds. Model point cloud (blue), target point cloud (pink) with outliers highlighted by cross markers.