

# Robust Computer Vision through Kernel Density Estimation

Haifeng Chen and Peter Meer

Electrical and Computer Engineering Department  
Rutgers University, Piscataway, NJ, 08854-8058, USA  
{haifeng, meer}@caip.rutgers.edu

**Abstract.** Two new techniques based on nonparametric estimation of probability densities are introduced which improve on the performance of equivalent robust methods currently employed in computer vision. The first technique draws from the projection pursuit paradigm in statistics, and carries out regression M-estimation with a weak dependence on the accuracy of the scale estimate. The second technique exploits the properties of the multivariate adaptive mean shift, and accomplishes the fusion of uncertain measurements arising from an unknown number of sources. As an example, the two techniques are extensively used in an algorithm for the recovery of multiple structures from heavily corrupted data.

## 1 Introduction

Visual data is complex and most often not all the measurements obey the same parametric model. For a satisfactory performance, robust estimators tolerating the presence of outliers in the data, must be used. These estimators are already popular in the vision community, see [17] for a representative sample of applications. Some of the robust techniques, like M-estimators and least median of squares (LMedS) were imported from statistics, while others, like Hough transform and RANSAC are innate, developed initially to solve specific vision problems.

It was shown in [2] that all the robust methods widely used in computer vision can be regarded as members of the same family of *M-estimators with auxiliary scale*, and that estimators with a smooth loss function (see Section 2) are to be preferred. In this paper we propose a novel approach toward computing M-estimators. The new approach combines several of the desirable characteristics of the different robust techniques already in use in computer vision and it is well suited for processing complex visual data.

A large class of computer vision problems can be modeled under the *linear errors-in-variables* (EIV) model. Under the EIV model *all* the measurements  $\mathbf{y}_i \in \mathcal{R}^p$  are corrupted independently by zero mean noise,  $\mathbf{y}_i = \mathbf{y}_{io} + \delta\mathbf{y}_i$ , where the subscript 'o' denotes the unknown true value. In the sequel will consider the simplest case in which the noise covariance is  $\sigma^2\mathbf{I}_p$ , however, our results can be easily extended to arbitrary covariance matrices. A linear constraint can be defined as

$$\mathbf{y}_{io}^\top \boldsymbol{\theta} - \alpha = 0 \quad i = 1, \dots, n \quad \|\boldsymbol{\theta}\| = 1. \quad (1)$$

When the constraint is written under this form it can be shown that the Euclidean distance between a measurement  $\mathbf{y}_i$  and  $\hat{\mathbf{y}}_i$ , its orthogonal projection on the hyperplane defined by the linear constraint is

$$\|\mathbf{y}_i - \hat{\mathbf{y}}_i\| = |\mathbf{y}_i^\top \boldsymbol{\theta} - \alpha| \quad (2)$$

i.e., the geometric distance and the algebraic distance are the same.

The optimal estimator for the above model is the *total least squares* (TLS) technique. In the presence of outliers the corresponding robust M-estimator is defined as

$$[\hat{\alpha}, \hat{\boldsymbol{\theta}}] = \underset{\alpha, \boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \rho \left( \frac{1}{s} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\| \right) \quad \text{subject to } \hat{\mathbf{y}}_i^\top \boldsymbol{\theta} - \hat{\alpha} = 0 \quad (3)$$

where  $s$  is a scale parameter, and  $\rho(u)$  is a nonnegative, even symmetric loss function, nondecreasing with  $|u|$  and having the unique minimum  $\rho(0) = 0$ . Only the class of *redescending* M-estimators is considered here, and therefore we can always assume that  $\rho(u) = 1$  for  $|u| > 1$ . A frequently used redescending M-estimator has the *biweight* loss function

$$\rho(u) = \begin{cases} 1 - (1 - u^2)^3 & \text{if } |u| \leq 1 \\ 1 & \text{if } |u| > 1 \end{cases} \quad (4)$$

and it will be the one employed throughout the paper.

The success of an M-estimation procedure, i.e., accurate estimation of the model parameters through rejection of the outliers, is contingent upon having a satisfactory scale parameter. (The issue of the breakdown point of the M-estimators is of lesser relevance in our context as will be seen later.) In [2] it was shown that the robust techniques imported from statistics differ from those developed by the vision community in the way the scale is obtained. For the former the scale is estimated from the data, while for the latter its value is set a priori.

It was implicitly assumed in (3) that a good scale estimate is already known. Statistical techniques for simultaneous estimation of the scale are available, e.g., [20], but they can not handle complex data of the type considered in this paper. Similarly, not in every vision application can a reliable scale estimate be obtained from the underlying physical properties. For example, in data containing multiple structures, i.e., several instances of the same model but with different sets of parameters, each structure may have been measured with a different uncertainty. A typical vision task generating such data is the recovery of the motion parameters for several moving objects in the presence of camera egomotion. In this case using a single global scale value for all the involved robust estimation procedures may not be satisfactory.

The performance of the robust M-type regression technique proposed in this paper has only a weak dependence on the accuracy of the scale estimate. Nevertheless, the technique provides a satisfactory inlier/outlier dichotomy for a wider range of contaminations than the traditional M-estimators. This performance improvement is achieved by recasting M-estimation as a kernel density estimation problem.

Kernel density estimation is a well known technique in statistics and pattern recognition. See the books [16] and [19] for a statistical treatment, and [7, Sec.4.3] for a pattern

recognition description. Let  $x_i, i = 1, \dots, n$ , be scalar measurements drawn from an arbitrary probability distribution  $f(x)$ . The kernel density estimate of this distribution  $\hat{f}(x)$  (called the Parzen window estimate in pattern recognition), is obtained based on a kernel function  $K(u)$  and a bandwidth  $h$  as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right). \tag{5}$$

The kernel functions considered here satisfy the following properties

$$\begin{aligned} K(u) = K(-u) \geq 0 & \quad K(0) \geq K(u) \quad \text{for } u \neq 0 \\ K(u) = 0 \quad \text{for } |u| > 1 & \quad \int_{-1}^1 K(u) = 1. \end{aligned} \tag{6}$$

Other conditions on the kernel function or on the density to be estimated [19, p.18], are always satisfied in practice. The even symmetry of the kernel function allows us to define its profile,  $k(u)$  from

$$K(u) = c_k k(u^2) \tag{7}$$

where  $c_k$  is a normalization constant determined by (6). The importance of the profile is revealed in the case of multivariate kernel density estimation which will be discussed in Section 3.

The quality of the density estimate  $\hat{f}(x)$  is assessed using the asymptotic mean integrated error (AMISE), i.e., the integrated mean square error between the true density and its estimate for  $n \rightarrow \infty$ , while  $h \rightarrow 0$  at a slower rate. The AMISE optimal bandwidth depends on the second derivative of  $f(x)$ , the unknown density [19, p.22]. While a good approximation of this bandwidth can be obtained employing a simple plug-in rule [19, p.72], for our purposes a bandwidth depending only on the kernel function and a raw scale estimate [19, Sec.3.2.2] suffices

$$\hat{h} = \left[ \frac{243R(K)}{35\mu_2(K)^2n} \right]^{1/5} \hat{\sigma} \quad \hat{\sigma} = c \operatorname{med}_j |x_j - \operatorname{med}_i x_i| \tag{8}$$

$$R(K) = \int_{-1}^1 K(u)^2 du \quad \mu_2(K) = \int_{-1}^1 u^2 K(u) du .$$

The data is taken into consideration through a median absolute deviations (MAD) scale estimate. The proportionality constant can be chosen as  $c = 0.5$  or  $1$  to avoid over-smoothing of the estimated density [19, p.62]. In Section 2 the connection between the regression M-estimator (3) and the univariate density estimation (5) is established and then exploited to computational advantage.

In Section 3 kernel density estimation is reformulated under its most general multivariate form and will provide the tool to solve the following difficult problem. Let  $\beta_j \in \mathcal{R}^p, j = 1, \dots, m$ , be a set of measurements whose uncertainty is also available through the covariance matrices  $\mathbf{C}_j$ . A large subset of these measurements is related

to  $M \ll m$  different data sources, while the rest are completely erroneous. The value of  $M$  is *not* known. Find the best (in statistical sense) estimates for the  $M$  vectors and covariances characterizing these sources. This is a fundamental feature space analysis problem and our approach, based on an extension of the variable bandwidth mean shift [4], provides a simple robust solution.

In Section 4 the two new techniques become the main building blocks of an algorithm for analyzing data containing multiple structures. The success of the algorithm is illustrated with 3D examples. Finally, in Section 5 the potential of the proposed techniques to solve difficult vision tasks, and the remaining open issues are discussed.

## 2 M-Estimators and Projection Pursuit

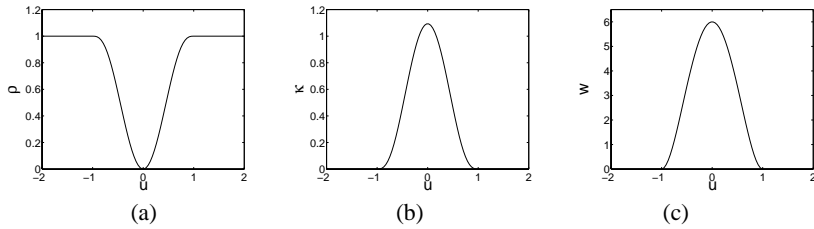
Projection pursuit is a nonparametric technique introduced in statistics in the 1970s for exploratory data analysis, and was soon applied also to nonlinear regression and density estimation problems. Projection pursuit seeks “interesting” low-dimensional (almost always one-dimensional) projections of multidimensional data. The informative value of a projection is measured with a *projection index*. Most often the projection index is a scalar functional of the univariate probability distribution estimated from the data projected on the chosen direction. The “best” direction is obtained through numerical optimization and corresponds to an extremum of the projection index. This is one step in an iterative procedure toward the solution. The current solution is then updated and a new search is initiated. For example, in projection pursuit regression at each iteration the shape of a smooth univariate nonlinear function which minimizes the sum of squared residuals is determined. At the subsequent iteration this function is incorporated into the current model. Convergence is declared when the squared error falls below a threshold. The papers [9], [11] offer not only excellent reviews on the projection pursuit paradigm, but also contain extensive discussions from researchers working on related topics.

There is a strong connection between robust linear regression estimators, such as least median of squares, and the projection pursuit procedure [6], [14, Sec.3.5]. This relationship, however, was investigated in statistics mostly for theoretical considerations and in the case of traditional regression, i.e., the case in which only one of the measured variables is corrupted by noise. In this paper will apply the projection pursuit paradigm to design a technique for the more general errors-in-variables (EIV) model which is better suited for vision tasks.

To prove the connection between EIV regression M-estimation and kernel density estimation, the definition (3) is rewritten as

$$[\hat{\alpha}, \hat{\boldsymbol{\theta}}] = \operatorname{argmax}_{\alpha, \boldsymbol{\theta}} \left[ 1 - \frac{1}{n} \sum_{i=1}^n \rho \left( \frac{\mathbf{y}_i^\top \boldsymbol{\theta} - \alpha}{s} \right) \right] = \operatorname{argmax}_{\alpha, \boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \kappa \left( \frac{\mathbf{y}_i^\top \boldsymbol{\theta} - \alpha}{s} \right) \quad (9)$$

where  $\kappa(u) = c_\rho [1 - \rho(u)]$  will be called the *M-kernel function*, and  $c_\rho$  is the normalization constant making  $\kappa(u)$  a proper kernel. Note that  $\kappa(u) = 0$  for  $|u| > 1$ , and that the even symmetry of the loss function  $\rho(u)$  allows the removal of the absolute values when (2) is plugged in. In Figs. 1a and 1b the biweight loss function and its corresponding M-kernel function is shown. Compare the M-kernel with the weight function



**Fig. 1.** The different functions associated with the *biweight* M-estimator. (a) The loss function  $\rho(u)$ . (b) The M-kernel function  $\kappa(u)$ . (c) The weight function  $w(u)$ .

$w(u) = u^{-1}\rho'(u)$  used in the well known iterative reweighted least squares implementation of M-estimators [13, p.306] (Fig. 1c). The two functions while look similar have different expressions.

Let  $\theta$  be a unit vector defining a line through the origin in  $\mathcal{R}^p$ . The projections of the data points  $\mathbf{y}_i$  on this line have the intrinsic coordinates  $x_i = \mathbf{y}_i^\top \theta$ . Given a kernel  $K(u)$  and the bandwidth  $\hat{h}$  (8), the estimated density of this sequence is

$$\hat{f}_\theta(x) = \frac{1}{n\hat{h}_\theta} \sum_{i=1}^n K\left(\frac{\mathbf{y}_i^\top \theta - x}{\hat{h}_\theta}\right) \tag{10}$$

where the dependence of the bandwidth on the direction of the projection (through the scale estimate of the projected points) was made explicit. The *mode* of the density estimate is defined as

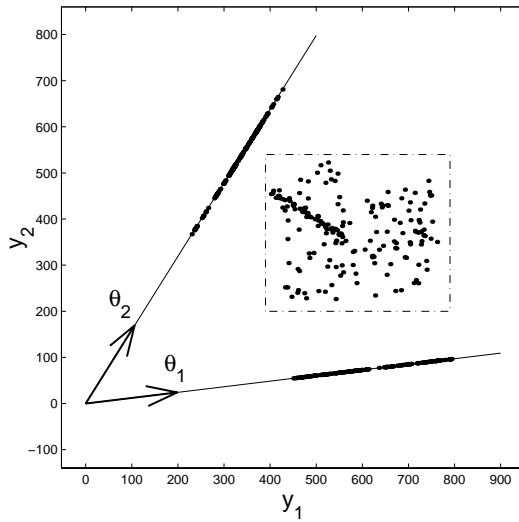
$$\hat{x}_\theta = \operatorname{argmax}_x \hat{f}_\theta(x) \tag{11}$$

and can be easily computed. Comparing (9) and (10) we can remark that if  $\kappa(u)$  is taken as the kernel function,  $\theta$  is chosen close to the true parameter of the linear model (1), and  $\hat{h}_\theta$  is a satisfactory substitute for the scale  $s$ , the mode (11) should provide a reasonable estimate for the intercept  $\alpha$ .

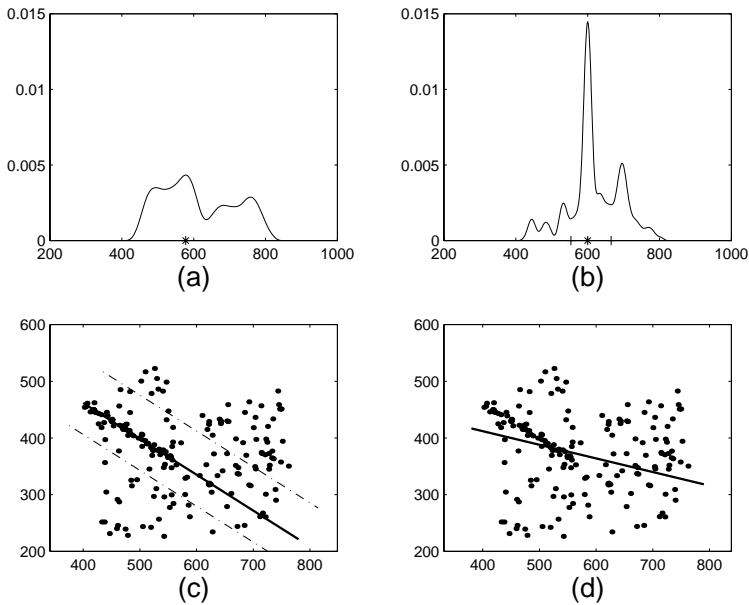
Based on the above observation the EIV linear model M-estimation problem can be reformulated as

$$[\hat{\alpha}, \hat{\theta}] = \operatorname{argmax}_\theta \left[ \hat{h}_\theta \max_x \hat{f}_\theta(x) \right] \tag{12}$$

which is the projection pursuit definition of the M-estimator, the projection index being the quantity inside the brackets. The equivalence between (3) and (12) is not perfectly rigorous since the scale was substituted with the bandwidth. However, this is an advantage since now the role of the scale is diminished and *any* bandwidth is satisfactory as long as it secures the reliable recovery of the mode. The bandwidth  $\hat{h}_\theta$  is proportional with the MAD scale estimate (8) which can be unreliable when the distribution is multimodal since the median is a biased estimator for nonsymmetric data. Similarly, for small measurement noise  $\hat{h}_\theta$  becomes small which can introduce artifacts if the bandwidth is not bounded downward.



**Fig. 2.** Projection pursuit principle: the parameter estimates are sought by examining the projections of the data points on arbitrary directions.



**Fig. 3.** Processing the data in Fig. 2. (a) Estimated density for the projection along direction  $\hat{\theta}_1$ . The detected mode is marked as \*. (b) Estimated density for the projection along direction  $\hat{\theta}_2$  which maximized the projection index. The points projecting inside the interval marked with the vertical bars are selected. (c) Projection pursuit based line estimate. The dashed lines bound the region delineated for robust postprocessing. (d) Hough transform based line estimate.

The projection pursuit approach toward M-estimation has a clear geometric interpretation. The direction  $\theta$  can be regarded as the unit normal of a candidate hyperplane fitted to the  $p$ -dimensional data,  $\mathbf{y}_i$ . The bandwidth  $\hat{h}_\theta$  defines a band centered on this plane. The band is translated in  $\mathcal{R}^p$  along  $\theta$  to maximize, *for the points within the band*, the weighted average of the orthogonal distances from the hyperplane. The M-estimate corresponds to the densest band (largest weighted average) over all  $\theta$ . Note the similarity with the well known interpretation of the LMedS estimator where the parameter estimates correspond to the narrowest band containing half the data points [14, p.126]. Our approach, however, has an important advantage. The optimization criterion is not dependent on a *preset* percentage of data points being inliers, thus yielding a better behavior in the presence of severely contaminated data, as it is shown in the following example.

The 180 data points in the rectangle in Fig. 2 belong to three classes. There are 50 measurements from the line segment  $0.54y_1 + 0.84y_2 - 606 = 0$  where  $400 \leq y_1 \leq 560$ , corrupted by normal noise with covariance  $5^2\mathbf{I}_2$ . A second structure is also present. Its 30 measurements are from the line segment  $0.54y_1 - 0.84y_2 - 60 = 0$  where  $600 \leq y_1 \leq 750$ , but were severely corrupted by normal noise with covariance  $20^2\mathbf{I}_2$  and became indistinguishable from the background. The background has 100 points uniformly distributed in the rectangle bounded by  $(425, 225)$  and  $(750, 525)$ . By definition the LMedS estimator cannot handle such data. Similarly, the global maximum of the Hough accumulator (built using all pairs of points) yields erroneous fits once the angle side of the bins exceeds 3.6 degrees. An example is shown in Fig. 3d.

The projections of the 2D data points on two directions are shown in Fig. 2. For the direction  $\theta_1 = [0.99, 0.12]$ , the computed bandwidth is  $h_{\theta_1} = 50.2$ . The mode is detected at  $\hat{x}_{\theta_1} = 578$  and has the value 0.004 (Fig. 3a). The projection index (12) is maximized by the direction  $\theta_2 = [0.52, 0.85]$ . The resulting bandwidth is  $h_{\theta_2} = 23.8$  and the mode at  $\hat{x}_{\theta_2} = 600$  has the value 0.013 (Fig. 3b).

The basin of attraction of the mode  $\hat{x}_{\theta_2}$ , is delineated by the first significant local minimum at the left and at the right, marked with vertical bars in Fig. 3b. They define two parallel lines in  $\mathcal{R}^2$  which bound the region containing the structure of interest (Fig. 3c). Since outliers relative to this structure may have been included, a robust postprocessing is required. The postprocessing also allows lower accuracy in the projection pursuit search for the best  $\theta$ , a necessary condition for searches in higher dimensional spaces (see Section 5).

We have used an M-estimator for robust postprocessing. The scale  $\hat{s}$  of the structure and its parameters  $\hat{\theta}$  were estimated simultaneously [13, p.307]. Finally, the inlier/outlier dichotomy is established and the robust covariance of the parameter estimate,  $\mathbf{C}_{\hat{\theta}}$ , is also computed. In the example, the final line estimate  $[0.53, 0.85, 604]$  is remarkable close to the true values in spite of the severe contamination (Fig. 3c). Here the improvement due to the postprocessing was small, however, its role is increased when the projection pursuit based M-estimator is employed as a computational module in Section 4.

### 3 Robust Data Fusion

The following problem appears under many forms in computer vision tasks. The  $m$  measurements  $\beta_j \in \mathcal{R}^p$  are available together with their uncertainty described by the covariance matrices  $\mathbf{C}_j$ . Taking into account these uncertainties, classify the measurements into  $M \ll m$  groups, where  $M$  is the number of clusters present in the data. The value of  $M$  is not known. The problem can also be regarded as a data fusion task in which the available evidence is to be reduced to the minimum number of plausible representations.

Will consider first the trivial case of  $M = 1$ , i.e., the case in which *all* the measurements belong to a single group. A satisfactory estimate for the center of the underlying cluster is obtained by minimizing the sum of Mahalanobis distances

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{j=1}^m (\beta - \beta_j)^\top \mathbf{C}_j^{-1} (\beta - \beta_j) \quad (13)$$

where the covariances are assumed to have full rank. As expected, the solution

$$\hat{\beta} = \left( \sum_{j=1}^m \mathbf{C}_j^{-1} \right)^{-1} \sum_{j=1}^m \mathbf{C}_j^{-1} \beta_j \quad (14)$$

is the covariance weighted average of the measurements. The more uncertain is a measurement (the inverse of its covariance has a smaller norm), the less it contributes to the result of the fusion.

To compute the covariance matrix  $\mathbf{C}$  (uncertainty) associated with  $\hat{\beta}$ , the covariances  $\mathbf{C}_j$  are approximated as  $\mathbf{C}_j \approx a_j \mathbf{C}$ . The common covariance structure  $\mathbf{C}$  and the positive proportionality factors  $a_j$  are determined from the minimization

$$[\hat{a}_j, \hat{\mathbf{C}}] = \operatorname{argmin}_{a_j, \mathbf{C}} \sum_{j=1}^m \|\mathbf{C}_j - a_j \mathbf{C}\|_F^2 \quad (15)$$

where  $\|\mathbf{B}\|_F^2 = \operatorname{trace}[\mathbf{B}^\top \mathbf{B}]$  is the squared Frobenius norm of the matrix  $\mathbf{B}$ . Differentiating after  $a_j$  and taking the matrix gradient after  $\mathbf{C}$ , two relations connecting the unknown quantities are obtained

$$\hat{\mathbf{C}} = \frac{\sum_{j=1}^m \hat{a}_j \mathbf{C}_j}{\sum_{j=1}^m \hat{a}_j^2} \quad \hat{a}_j = \frac{\operatorname{trace}[\mathbf{C}_j^\top \hat{\mathbf{C}}]}{\operatorname{trace}[\hat{\mathbf{C}}^\top \hat{\mathbf{C}}]} \quad (16)$$

The relations are evaluated iteratively starting from all  $\hat{a}_j = 1$ , which makes  $\hat{\mathbf{C}}$  the average covariance. The  $\hat{a}_j$ -s are then refined, and the next value of  $\hat{\mathbf{C}}$  is the one retained.

Will return now to kernel density estimation. A radially symmetric,  $p$ -dimensional multivariate kernel  $K(\mathbf{u})$  is built from the profile  $k(u)$  as

$$K(\mathbf{u}) = c_{k,p} k(\mathbf{u}^\top \mathbf{u}) \quad (17)$$

where  $c_{k,p}$  is the corresponding normalization constant and  $\mathbf{u} \in \mathcal{R}^p$ . The properties (6) can be easily extended to  $\mathcal{R}^p$ . In the most general case the bandwidth  $h$  is replaced by a symmetric positive definite *bandwidth matrix*,  $\mathbf{H}$ .

Given the data points  $\mathbf{x}_i, i = 1, \dots, n$ , in  $\mathcal{R}^p$ , their multivariate density estimate computed with the kernel  $K(\mathbf{u})$  and the bandwidth matrix  $\mathbf{H}$  is [16, Sec.4.2.1]

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) \tag{18}$$

$$K_{\mathbf{H}}(\mathbf{x}) = [\det[\mathbf{H}]]^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}) = c_{k,p} [\det[\mathbf{H}]]^{-1/2} k(\mathbf{x}^\top \mathbf{H}^{-1} \mathbf{x}). \tag{19}$$

Note that  $\mathbf{H} = h^2 \mathbf{I}_p$  reduces (19) to the well known, traditional multivariate kernel density estimation expression.

In practice using a single bandwidth is often not satisfactory since the available data points are not spread uniformly over the region of existence of the unknown density. The *sample point* kernel density estimator is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}_i}(\mathbf{x} - \mathbf{x}_i) \tag{20}$$

where each data point  $\mathbf{x}_i$  is considered in the computations through its own bandwidth matrix  $\mathbf{H}_i$ . The sample point estimator has superior performance relative to kernel density estimators in which the variable bandwidth is associated with the center of the kernel  $\mathbf{x}$ , [16, Sec.5.3]. From (20), taking into account (19) we obtain

$$\hat{f}(\mathbf{x}) = \frac{c_{k,p}}{n} \sum_{i=1}^n [\det[\mathbf{H}_i]]^{-1/2} k((\mathbf{x} - \mathbf{x}_i)^\top \mathbf{H}_i^{-1} (\mathbf{x} - \mathbf{x}_i)). \tag{21}$$

To solve the robust data fusion problem will compute the sample point density estimate of the  $m$  measurements  $\beta_j$ . Multivariate Epanechnikov kernels built from the profile [19, p.30]

$$k(u) = \begin{cases} 1 - u & 0 \leq u \leq 1 \\ 0 & u > 1 \end{cases} \tag{22}$$

are used, and as bandwidth matrices the covariances  $\mathbf{C}_j$  are employed. The covariance matrices are scaled to  $\chi_{\gamma,p}^2 \mathbf{C}_j$ , where  $\chi_{\gamma,p}^2$  is the chi-square value for  $p$  degrees of freedom and level of confidence  $\gamma$  (in our implementation  $\gamma = 0.995$ ). Thus

$$K_{\mathbf{C}_j}(\mathbf{u}) = 0 \quad \text{for} \quad \mathbf{u}^\top \mathbf{C}_j^{-1} \mathbf{u} > \chi_{\gamma,p}^2 \quad j = 1, \dots, m \tag{23}$$

i.e., the kernel associated with a measurement is nonzero in the region of confidence of that measurement having coverage probability  $\gamma$ . The density estimate (21) becomes

$$\hat{f}(\beta) = \frac{c_{k,p}}{m [\chi_{\gamma,p}^2]^{p/2}} \sum_{j=1}^m [\det[\mathbf{C}_j]]^{-1/2} k\left(\frac{1}{\chi_{\gamma,p}^2} (\beta - \beta_j)^\top \mathbf{C}_j^{-1} (\beta - \beta_j)\right). \tag{24}$$

Taking into account (22) we have obtained that solving the minimization problem (13) is equivalent to finding the maximum of the density estimate (24), i.e., its mode. (The apparent differences are only scalar normalization factors for the covariances.)

We are now ready to proceed to the proposed problem where the measurements come from an unknown number of sources  $M$ . To characterize these sources, first the  $M$  clusters are to be delineated, which as will be shown below is equivalent to finding *all* the significant modes of the density  $\hat{f}(\beta)$

$$\hat{\beta}_l = \arg \max_{\beta} \hat{f}(\beta) \quad l = 1, \dots, M. \quad (25)$$

Note that the value of  $M$  is determined automatically from the data. A mode of  $\hat{f}(\beta)$  corresponds to a zero of its gradient

$$\begin{aligned} \nabla \hat{f}(\beta) = \frac{c_{k,p}}{m [\chi_{\gamma,p}^2]^{(p/2+1)}} \sum_{j=1}^m [\det[\mathbf{C}_j]]^{-1/2} \mathbf{C}_j^{-1} (\beta - \beta_j) \times \\ \times k' \left( \frac{1}{\chi_{\gamma,p}^2} (\beta - \beta_j)^\top \mathbf{C}_j^{-1} (\beta - \beta_j) \right). \end{aligned} \quad (26)$$

The function  $g(u) = -k'(u)$  defines a new profile which in our case is

$$I_\gamma(\mathbf{u}) = g \left( \frac{\mathbf{u}^\top \mathbf{C}_j^{-1} \mathbf{u}}{\chi_{\gamma,p}^2} \right) = \begin{cases} 1 & \mathbf{u}^\top \mathbf{C}_j^{-1} \mathbf{u} \leq \chi_{\gamma,p}^2 \\ 0 & \mathbf{u}^\top \mathbf{C}_j^{-1} \mathbf{u} > \chi_{\gamma,p}^2 \end{cases} \quad (27)$$

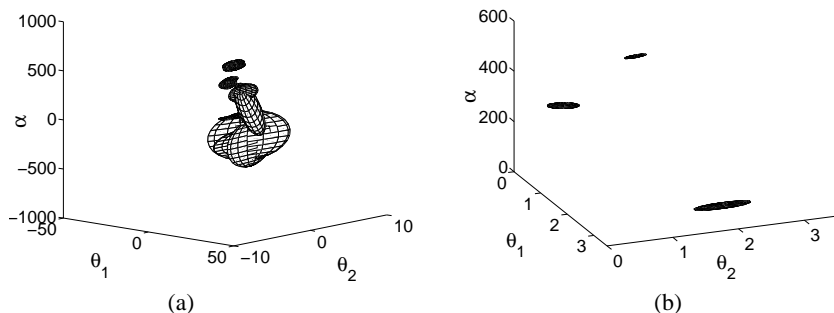
i.e., the indicator function selecting the data points inside the region of confidence of  $\beta_j$ . Defining the matrix  $\mathbf{W}_j = [\det[\mathbf{C}_j]]^{1/2} \mathbf{C}_j$  the expression of the gradient (26) can be rewritten

$$\begin{aligned} \nabla \hat{f}(\beta) = \frac{c_{k,p}}{m [\chi_{\gamma,p}^2]^{(p/2+1)}} \left( \sum_{j=1}^m I_\gamma(\beta - \beta_j) \mathbf{W}_j^{-1} \right) \times \\ \times \left[ \left( \sum_{j=1}^m I_\gamma(\beta - \beta_j) \mathbf{W}_j^{-1} \right)^{-1} \left( \sum_{j=1}^m I_\gamma(\beta - \beta_j) \mathbf{W}_j^{-1} \beta_j \right) - \beta \right] \end{aligned} \quad (28)$$

where the presence of the indicator function assures the robustness computations. Indeed, the zeros of the gradient are given by an expression similar to (14), but with the computations restricted to local regions in  $\mathcal{R}^p$ . As long as the  $M$  clusters are reasonably separated, computing their centers is based only on the appropriate data points. By choosing a kernel other than Epanechnikov from the beta family [19, p.31], instead of a binary indicator function (27) additional weighting can be introduced in (28).

The modes of  $\hat{f}(\beta)$  by definition are located in high density regions of  $\mathcal{R}^p$ . A versatile, robust mode detector is based on the *mean shift* property introduced first in pattern recognition [8], and which recently became popular in computer vision for a large variety of tasks [3]. The variable bandwidth version of the mean shift procedure was also developed [4].

The mean shift procedure recursively evaluates the second term of (28). The procedure starts by taking  $\beta = \beta_j$ , and a new value of  $\beta$  is computed using only the data



**Fig. 4.** An example of data fusion for  $p = 3$ ,  $m = 60$  and  $M = 3$ . (a) The measurements with regions of confidence. (b) The result of the multivariate variable bandwidth mean shift. Note the smaller scale.

points which yield nonzero values for the indicator function. The process is then repeated with the obtained  $\beta$ , i.e., the kernels are shifted according to the result of the previous step. Convergence is achieved when the shift becomes less than a threshold. See [3] and [4] for details about the mean shift procedure.

After the mean shift procedure was applied to all the  $m$  measurements, by associating these measurements with their point of convergence, arbitrarily shaped *basins of attraction* can be defined. Note that outliers, i.e., isolated erroneous measurements are not taken into account since they will fail to evolve. The points of convergence are characterized applying (14) and (16) to the data points in the basin of attraction. Pairs whose squared Mahalanobis distance is less than  $\chi_{\gamma,p}^2$  (under both metrics) are merged. The resulting  $M$  modes are the output of the robust fusion procedure. An example is shown in Fig. 4. The large confidence regions in Fig. 4a correspond to erroneous measurements and hide the majority of the data. After robust fusion three modes are detected, each associated with a small uncertainty (Fig. 4b).

The fusion technique introduced here can provide a robust component for more traditional approaches toward combining classifiers, e.g., [18], or for machine learning algorithms which improve performance through resampling, e.g., bagging [1].

## 4 Robust Regression for Data with Multiple Structures

Data containing multiple structures is characterized by the presence of several instances of the same model, in our case (1), each defined with a different set of parameters. The need for reliable processing of such data distinguishes estimation problems in computer vision from those in applied statistics.

The assumption that the sought model is carried by the absolute majority of the data points, is embedded in all robust estimators in statistics. In vision tasks, such as, structure from motion, 3D scene representation, this assumption is violated once information about more than one object is to be acquired simultaneously. Among the four main classes of robust techniques employed in vision (see Section 1) only the Hough transform has

the capability to handle complex multiple structured data. However, as our example in Section 2 has already shown, good performance of the Hough transform is contingent upon having access to the correct scale estimate (accumulator bin size), which in practice is often not possible. See [2] for a detailed discussion on the difficulties of traditional robust techniques in handling multiple structured data.

Four main processing steps can be distinguished in the implementation of the robust estimators based on a nondifferentiable optimization criterion: LMedS, RANSAC and Hough transform. First, several small random subsets of data points, i.e., samples, are selected. Next, from each sample a parameter estimate candidate is computed. In the third step, the quality of the candidates is assessed using all the data points and the candidate yielding the “best” quality measure is retained. Finally, the data is classified into inliers and outliers in relation to the model parameter estimates.

While some of these four steps can be intertwined and refined (or in the case of Hough transform disguised), they provide a general processing principle. This principle is still obeyed when the two techniques introduced in the paper are employed as computational modules in an algorithm for analyzing data with multiple structures.

1. *Definition of the random samples.*

The data is quantized in  $\mathcal{R}^p$  by defining a  $p$ -dimensional *bin* using the bandwidths (8) computed with a uniform kernel separately for each coordinate. The bins are ranked by the number of points inside, and at random one is chosen from the upper half of the ranking. Starting from this bin a *sample* is generated by probabilistic region growing. Any bin at the boundary of the current region selects a neighbor not yet in the region with probability equal to the normalized number of points of the neighbor. Normalization is by the total number of points of such neighbors. Region growing stops when the sample reaches the upper bound of allowed bins (in our 3D examples 6% of all nonempty bins), or no further growing is possible.

2. *Computation of the parameter estimate candidates.*

For each of  $N$  samples (60 in our experiments) the projection pursuit based M-estimation procedure discussed in Section 2 is applied. For each sample the candidate vector  $\hat{\theta}_l$  its covariance  $\mathbf{C}_{\hat{\theta}_l}$ , and a scale estimate  $\hat{s}_l$  are obtained. For display purposes, the points declared inliers are delineated with a *bounding box*.

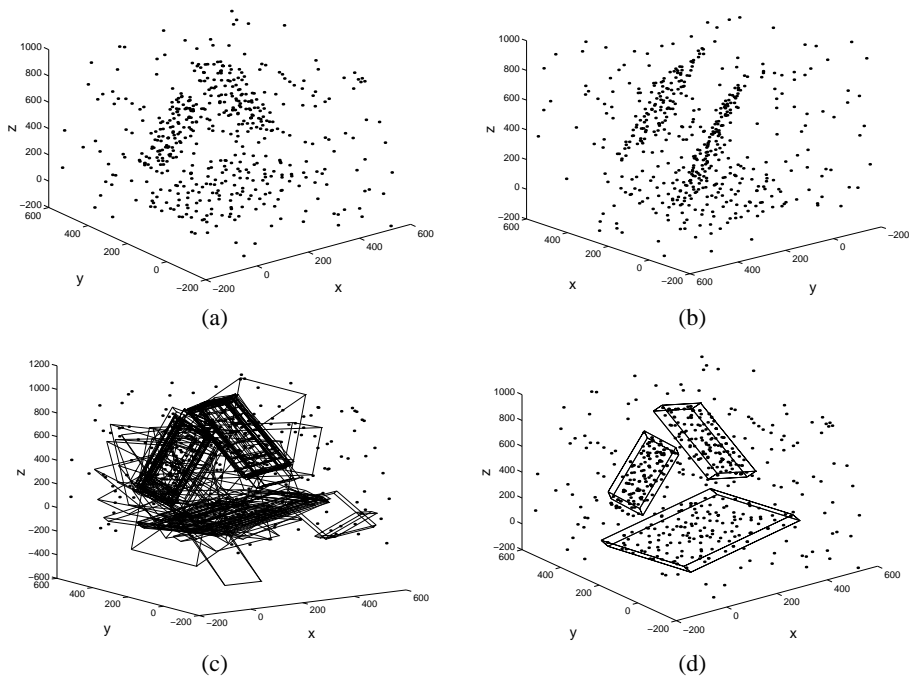
3. *Selection of the best candidates.*

Using the  $N$  estimates and their covariances, the robust fusion procedure discussed in Section 3 is applied. The number of structures  $M$  present in the data is determined and their characteristics are computed.

4. *Classification of the data.*

To refine the relation between the  $M$  structures and the data points declared inliers in the samples, each sample/structure association receives a vote. Only the points with more than 4 votes are retained for a structure. Finally, starting with the structure having the largest number of points, they are recursively removed from the data. Since the data classification starts from a reliable basis, other more sophisticated or application specific procedures can also be used.

Two experiments with 3D synthetic data containing  $M = 3$  structures, are presented here. The first data set (Figures 5a and 5b) contains three planar regions in a chevron-type arrangement. Each region contains 100 points corrupted with normal noise having



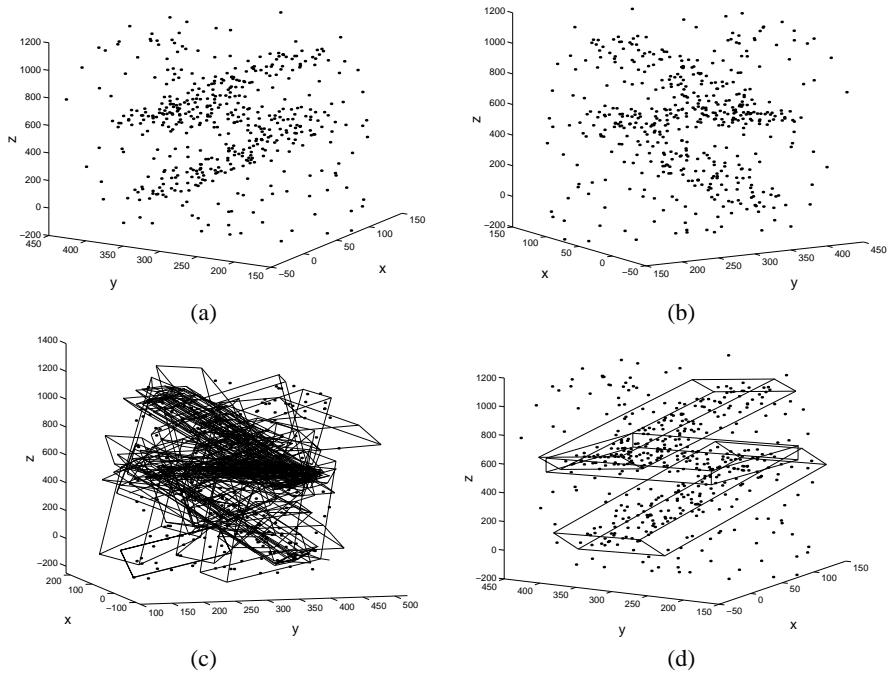
**Fig. 5.** Example of 3D data analysis containing multiple structures. (a), (b) Two views of the data. (c) Bounding boxes of the employed  $N = 60$  samples. (d) Delineated structures.

covariance  $10^2 \mathbf{I}_3$ . In the background 200 more data points are scattered uniformly in a cube incorporating all three structures. The 60 bounding boxes resulting at the end of the M-estimation procedures are shown in Figure 5c, while the feature space and the result of the robust fusion are in Figure 4. The output of the algorithm, the three structures delineated by their final bounding boxes, is shown in Figure 5d.

The second data set (Figures 6a and 6b) has the same characteristics, however, the three planar regions are now arranged in a Z-type configuration. In spite of the intersecting regions, the algorithm succeeded to distinguish the structures (Figure 5d). In both examples the estimated parameters were close to the true values for the planes.

## 5 Discussion

Many computer vision problems can be recast under the framework of robust analysis (regression) of data containing multiple structures. For example, the Costeira-Kanade algorithm for structure-from-motion factorization for multiple objects [5], was recently reformulated by Kanatani as finding for each tracked object a four-dimensional linear subspace in a space having the dimension twice the number of image frames [12]. Similarly, to build from an image sequence a scene-based representation of the visual environment, e.g. a mosaic, the multiple layer plane+parallax representation is the most



**Fig. 6.** Example of 3D data analysis containing multiple structures. (a), (b) Two views of the data. (c) Bounding boxes of the employed  $N = 60$  samples. (d) Delineated structures.

general model [10], which can be also used for detecting independently moving objects [15]. The algorithm proposed in this paper offers a tool which can simultaneously extract all the significant model instances, instead of the usually employed recursive approach in which the “dominant” feature is detected first.

These vision tasks, however, require processing in high dimensional spaces. Thus, an efficient search strategy over  $\theta$  has to be employed when the projection index is maximized (M-estimation). In the 3D examples described above, first  $4^2$  directions distributed uniformly over  $\mathcal{R}^3$  were used, followed by a refinement of another  $4^2$  around the “best” direction from the previous step. The 3D examples were processed in MATLAB in less than a minute. Using a parametrization which takes into account that  $\theta$  is a unit vector [20], we are currently developing a computationally feasible search strategy for higher dimensions. Ideally the search should also take into account a priori information specific to the vision task to be solved.

The two techniques presented in the paper make extensive use of nonparametric statistics tools which are more sensitive than the parametric methods, however, require more supporting data points to yield reliable results. See for example, [9, p.473] for a discussion of projection pursuit for small sample sizes. Nevertheless, the new data analysis algorithm tolerates “bad” data better than the robust techniques traditionally employed in computer vision.

**Acknowledgment.** The support of the NSF grant IRI 99-87695 is gratefully acknowledged.

## References

1. L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.
2. H. Chen, P. Meer, and D. E. Tyler. Robust regression for data with multiple structures. In *2001 IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 1069–1075, Kauai, HI, December 2001.
3. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24, May 2002.
4. D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *8th International Conference on Computer Vision*, volume I, pages 438–445, Vancouver, Canada, July 2001.
5. J. Costeira and T. Kanade. A multiple factorization method for motion analysis. *International J. of Computer Vision*, 29:159–179, 1998.
6. D. Donoho, I. Johnstone, P. Rousseeuw, and W. Stahel. Discussion: Projection pursuit. *Annals of Statistics*, 13:496–500, 1985.
7. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, second edition, 2000.
8. K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 1975.
9. P. J. Huber. Projection pursuit (with discussion). *Annals of Statistics*, 13:435–525, 1985.
10. M. Irani and P. Anandan. Video indexing based on mosaic representations. *Proceedings of IEEE*, 86:905–921, 1998.
11. M. C. Jones and R. Sibson. What is projection pursuit? (with discussion). *J. of the Royal Stat. Soc. Series A*, 150:1–37, 1987.
12. K. Kanatani. Motion segmentation by subspace separation and model selection. In *8th International Conference on Computer Vision*, volume II, pages 301–306, Vancouver, Canada, July 2001.
13. G. Li. Robust regression. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Exploring Data Tables, Trends, and Shapes*, pages 281–343. John Wiley & Sons, 1985.
14. P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.
15. H. S. Sawhney, Y. Guo, and R. Kumar. Independent motion detection in 3D scenes. *IEEE Trans. Pattern Anal. Machine Intell.*, 22:1191–1199, 2000.
16. B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.
17. Special Issue. Robust statistical techniques in image understanding. *Computer Vision and Image Understanding*, 78, April 2000.
18. D. M. J. Tax, M. van Breukelen, R. P. W. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recog.*, 33:1475–1486, 2000.
19. M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.
20. R. H. Zamar. Robust estimation in the errors-in-variables model. *Biometrika*, 76:149–160, 1989.