Model Selection Techniques and Merging Rules for Range Data Segmentation Algorithms

Kishore Bubna   Charles V. Stewart

Department of Computer Science,
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
{bubnak, stewart}@cs.rpi.edu

Running head: Model Selection Techniques

Correspondence:

Charles V. Stewart
Associate Professor
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180
Email: stewart@cs.rpi.edu
Phone: (518) 276 6731
Fax: (518) 276 4033

Abstract

The problem of model selection is relevant to many areas of computer vision. Model selection criteria have been used in the vision literature and many more have been proposed in statistics, but the relative strengths of these criteria have not been analyzed in vision. More importantly, suitable extensions to these criteria must be made to solve problems unique to computer vision. Using the problem of surface reconstruction as our context, we analyze existing criteria using simulations and sensor data, introduce new criteria from statistics, develop novel criteria capable of handling unknown error distributions and outliers, and extend model selection criteria to apply to the surface merging problem. The new and existing model selection criteria and merging rules are tested over a wide range of experimental conditions using both synthetic and sensor data. The new surface merging rules improve upon previous results, and work well even at small step heights ($h = 2\sigma$) and crease discontinuities. Our results show that a Bayesian criteria and its bootstrapped variant perform the best, although for time-sensitive applications, a variant of the Akaike criterion may be a better choice. Unfortunately, none of the criteria work reliably for small region sizes, implying that model selection and surface merging should be avoided unless the region size is sufficiently large.

Machine vision systems extract useful information from images in order to perform specific tasks. Estimating a geometric model forms the basis of this extraction process. While some physical processes are well understood and easy to model mathematically, in most cases different models must be fit to the data and the best model is selected from these competing models. This process, generally referred as *model selection*, must precede parameter estimation when the model is not known a priori. It arises in diverse machine vision problems. For example, the best camera calibration model must be selected to get unbiased data from sensors, the correct deformation model must be selected to describe deviations from CAD specifications when inspecting manufactured parts, and surfaces must be defined using the correct mathematical model in surface reconstruction for reverse engineering and 3D modeling. While the problem of parameter estimation is well studied in computer vision [8, 9, 33, 34, 42, 43], the associated problem of model selection has only received recent attention in the literature [13, 49]. Yet without good solutions to the model selection problem, the estimated parameters have little meaning.

Model selection criteria in vision have different origins. Many of these criteria are heuristics [6, 8, 39, 47] and some rely on user defined thresholds. Others, especially recent ones, are applications of statistical and information theoretic criteria [7, 9, 17, 28, 29, 30, 45, 52, 53]. Unfortunately, the advantages and limitations of these criteria in computer vision algorithms have not been carefully analyzed. Most do not work well for small region sizes, many make errors near small magnitude discontinuities, and some are biased towards higher or lower order models. Further, model selection criteria in vision must tolerate outliers [9, 17], unknown noise distributions, and other kinds of unmodeled errors in the data [16].

Model selection criteria have been derived to choose a single model — e.g. a planar, quadratic, or higher order model — for a given set of data, but problems such as surface merging require criteria that can decide between describing a data set with a single model or partitioning the data set and describing each with a separate model (see fig. 1). Merging techniques used in vision are based on empirical heuristics [6, 15, 48, 39], and perform poorly at small discontinuities. Further, in an attempt to avoid the model selection problem, many merging techniques only join fits to the same model [6, 15, 27, 48], potentially limiting the effectiveness of merging. Hence, mathematical criteria to merge regions and to simultaneously decide the correct model for a merged region must be formulated.

One computer vision problem where model selection and merging techniques are crucial is surface reconstruction. Many reconstruction algorithms use a local-to-global approach in which parameter estimation techniques and local decision criteria are combined in a greedy surface recovery strategy. This approach involves estimating initial surface patches (using grid techniques [15, 48], clustering methods [23, 39], or by region growing [6, 9, 29, 47]), and later pruning redundant surface patches [9, 29], or merging artificial surface boundaries [6, 15, 39, 48]. In the absence of a priori information, model selection forms an important part of each step. For example, when expanding "seed regions", at each iteration it must be decided whether to continue growing using the same model or to switch to a different model. When pruning redundant fits a model selection criteria may be used explicitly [9] or combined with greedy search techniques [29]. When merging adjacent surfaces, a criterion must be used to determine if the data should be represented by a single fit or by two or more different fits.

Surface reconstruction, therefore, provides a good context for studying the model selection and merging problems in computer vision. Using this problem as our context, we study the characteristics of different model selection criteria. We modify them for use in the presence of outliers, and develop new criteria based on bootstrapped data distributions [19] which do not require a prior model of the noise distribution. Finally, we extend model selection criteria to develop new techniques for surface merging. All new and existing criteria studied in this paper are free from user-defined, data dependent, thresholds, although some use statistical thresholds (confidence intervals). We compare the relative performance of these criteria using simulated data (containing small-scale Gaussian errors), and real data (containing small-scale random errors and outliers). Our results show that these new criteria may be used to give improved performance over existing techniques (for example, the discontinuity in Figure 2 can be detected by using new techniques presented in this paper). The experiments on simulated and sensor data determine the performance of these criteria under different conditions, and identify situations in which they perform poorly. These results, therefore, may be used to decide among different model selection criteria and merging criteria for different types of data and applications.

1 Definitions

Range image: A range image is characterized by a point $\boldsymbol{p}_i = [\boldsymbol{x}_i \; z_i]^T$ at any pixel i in the image. For our simulations, $\boldsymbol{x}_i$ will simply be a scalar x_i, and for real range images $\boldsymbol{x}_i = [x_i \; y_i]^T$. We call the former 2D range images and the latter 3D range images. For this paper, we assume errors in range data are all in the depth (z) direction[1].

[1]Errors in sensor data are, generally, along all coordinate directions. But, experiments with our sensors show that for relatively small fields of view (a viewing cone of 25 degrees or less), the errors can be approximated to be along the depth (z) direction. Almost all other algorithms assume the same [2], and have not reported any problems.

Candidate models and parameter estimation: Experiments in this paper are based on data sets from linear and quadratic models. To test performance of different criteria we use the set $M = \{m_0, m_1, m_2, m_3\}$ of candidate models, where m_i stands for the ith order model. Models m_0 and m_3 are included in M to detect bias toward low or high order models in different criteria. The models in M use discrete orthogonal polynomials as basis functions [3, 5], and are given by

$$z(\boldsymbol{x}) = \sum_{j=0}^{d_m-1} \theta_j \; \phi_j(\boldsymbol{x}), \quad m = 0 \ldots 3, \tag{1}$$

where d_m is the number of parameters in the model. Orthogonal polynomials are used because they give well-conditioned matrices, and estimation is efficient because fit to high order models builds on fits to lower order models (the second advantage is lost, however, when using robust techniques because outliers are determined differently for different models, and parameters must be estimated separately). The parameter vector is given by $\boldsymbol{\theta}_m = [\theta_0 \; \theta_1 \; \ldots \theta_{d_m-1}]^T$. The set of orthogonal basis polynomials, $\phi_j(\boldsymbol{x})$, is constructed using the n data points, and satisfies the relation [3, 12]

$$\sum_{i=1}^{n} \phi_p(\boldsymbol{x}_i) \; \phi_q(\boldsymbol{x}_j) = 0, \quad \text{for } p \neq q. \tag{2}$$

In this paper, we consider models of the form

$$\boldsymbol{Z} = \boldsymbol{X}_m \; \boldsymbol{\theta}_m \; + \; \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{Z}$ contains the $(n \times 1)$ depth values, $\boldsymbol{X}_m$ contains $(n \times d_m)$ orthonormal polynomials where any element $\boldsymbol{X}_m(i,j) = \phi_j(\boldsymbol{x}_i)$, and $\boldsymbol{e} = [e_1 \; e_2 \; \ldots e_n]^T$ is a vector of unobserved, but independent random variables. Note that the standard deviation of noise, σ, may or may not be known *a priori*.

Estimates of $\boldsymbol{\theta}_m$ and σ, obtained by fitting model m to the data, are denoted by $\hat{\boldsymbol{\theta}}_m$ and $\hat{\sigma}_m$, respectively. Information-theoretic criteria use the loglikelihood of estimated parameters for model selection, hence, maximum likelihood estimators (MLEs) must be used for parameter estimation.

We use ordinary least-squares for data with Gaussian errors, and following [9], we use iteratively reweighted least squares (IRLS) [24] with an M-estimator based on t-distribution for data with outliers. In the latter case, IRLS is initialized using least median of squares (LMS) [33]. We denote the likelihood for model m by $L(\boldsymbol{\theta}_m)$ and the residual sum of squares by RSS_m. The covariance matrix for $\boldsymbol{\theta}_m$ is given by $\boldsymbol{V}(\boldsymbol{\theta}_m)$.

2 Intuition about model selection

It is well known and easily demonstrated that a higher order model fits any dataset more accurately than a lower order model. Thus, accuracy as a sole measure of fit quality is ineffective when comparing best fits from different models; fit accuracy must be combined with other fit characteristics in order to choose the correct model. Consider model selection for noisy data points $(\{(1, 0.8), (2, 2.1), (3, 2.9), (4, 3.8)\})$ from the straight line $z = x$. Figure 3(a) shows the zeroth-order (m_0), first-order (m_1), and second-order (m_2) fits to the data. Observe how the quadratic model fits the data best, although the linear model is the correct model. Now consider another sampling of the same line given by $(\{(1,\ 1.1),\ (2,\ 1.8),\ (3,\ 2.8),\ (4,\ 4.1)\})$. Figure 3(b) shows the fits to this new set of points. In this case, the m_0 and m_1 fits remain almost the same as in fig. 3(a), but the quadratic fit changes significantly and flips to the other side of the linear fit. In this situation, a linear model is more "stable" than the quadratic and therefore intuitively appears to be preferable despite being slightly less accurate. An overly accurate fit also models part of the random noise which it is supposed to remove, making the estimated parameters very sensitive to different samplings of the same data points.

While fit accuracy has measures such as residual sum of squares and likelihood at the estimated parameters (see sec. 1), measures of model stability are less well-known. For a model to be stable, its estimated parameters, say, $\hat{\boldsymbol{\theta}}_m$ and $\hat{\boldsymbol{\theta}}_m^{'}$ from two different samplings must be close to each other. Thus, a model that gives a more "compact" set of $\hat{\boldsymbol{\theta}}_m$s due to slight perturbations in the data is likely to be more stable. In [12], we show that this "compactness" can be measured by the covariance matrix, $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$, of the estimated parameter vector $\hat{\boldsymbol{\theta}}_m$. In fact, the stability of a model turns out to be directly proportional to $|\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)|^{1/2}$. Note that this measure of stability does not say that simple models are stabler. But simpler models indeed have a higher value of $|\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)|^{1/2}$. Also, two models with the same number of parameters are not treated equally; the model with a higher value of $|\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)|^{1/2}$ has more stability. This measure of fit stability arises in several information theoretic criteria in sec. 3.1.

3 Model selection

This section gives an overview of model selection criteria in the literature. While some of these criteria have previously been used in surface segmentation algorithms, others are new to computer vision and have been borrowed from the statistics literature. The section first introduces the information-theoretic model selection criteria and then discusses the model selection criteria based on hypothesis tests.

3.1 Information-theoretic criteria

This section discusses model selection criteria based on Bayes rule, Kullback-Leibler (K-L) distance, and minimum description lengths (MDL), and shows how the intuition discussed in sec. 2 ties with different criteria. In each case, we give a brief description of the basic principles, discuss the assumptions used, and present the criterion for the case when σ is known and when it is not known *a priori*.

3.1.1 Model selection using Bayes rule

Criteria based on Bayes rule choose the model that maximizes the probability of the data, D, given the model m and prior information I. This probability is denoted by $P(D|m, I)$. Using Bayes rule (and assuming the parameter vector, $\boldsymbol{\theta}_m$, and standard deviation of noise, σ, are independent [32, page 109]),

$$P(D|m, I) = \int \int P(D|\boldsymbol{\theta}_m, \sigma, m, I)\ P(\boldsymbol{\theta}_m|m, I) P(\sigma|I)\ d\boldsymbol{\theta}_m d\sigma \tag{4}$$

$P(D|\boldsymbol{\theta}_m, \sigma, m, I)$ in (4) is just the likelihood $L(\boldsymbol{\theta}_m)$. $P(\boldsymbol{\theta}_m|m, I)$ is the prior probability of $\boldsymbol{\theta}_m$. Since reconstruction applications generally lack prior information on parameters, we use a uniform prior on $\boldsymbol{\theta}_m$ (see [22, appendix A]). When σ is known, its prior, $P(\sigma|I)$, is a delta function at the known σ, and (4) reduces to an integral with respect to $\boldsymbol{\theta}_m$ only. Solving this reduced integral using a second order Taylor's expansion of $\log L(\boldsymbol{\theta}_m)$ at $\hat{\boldsymbol{\theta}}_m$ [26, chapter 24] yields

$$P(D|m, I) \approx (2\pi)^{d_m/2} L(\hat{\boldsymbol{\theta}}_m)[|\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m)|]^{1/2}, \tag{5}$$

Notice how the accuracy term given by $L(\hat{\boldsymbol{\theta}}_m)$ and the stability term given by $|\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m)|$ are combined in this criteria.

When σ is not known, we need to assign $P(\sigma|I)$. Again, we use non-informative priors on $P(\sigma|I)$. For the Gaussian case, using the non-informative prior $1/\sigma$ for σ (see [26, chapter 6, page 29]), and assigning other probabilities as before, (4) reduces to

$$P(D|m,I) = \frac{\Gamma((n-d_m)/2)}{2^{(d_m/2)+1}\ \pi^{n/2}\ |\boldsymbol{X}_m^T\boldsymbol{X}_m|^{1/2}\ RSS_m^{(n-d_m)/2}}, \tag{6}$$

where $\Gamma(\cdot)$ is the standard Gamma function, and RSS_m is the residual sum of squares for model m. Alternatively, assuming a uniform prior on σ [22], (4) reduces to

$$P(D|m,I) = (2\pi)^{d_m/2}\ L(\hat{\boldsymbol{\theta}}_m,\hat{\sigma})\ [|\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m,\hat{\sigma})|]^{-1/2}, \tag{7}$$

These criteria, (5), (6) and (7), will be referred to as BAYES.

To avoid the expense of estimating $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$, several asymptotic approximations of (5) have been introduced. A common one, due to Schwarz [41] is given by

$$P(D|m,I) \approx L(\hat{\boldsymbol{\theta}}_m)\ n^{-d_m/2}, \tag{8}$$

and is commonly known as BIC. Once again $L(\hat{\boldsymbol{\theta}}_m,\hat{\sigma}_m)$ replaces $L(\hat{\boldsymbol{\theta}}_m)$ in (8) when σ is unknown.

3.1.2 Model selection using K-L distance

Some of the earliest criteria select the model minimizing the Kullback-Leibler (K-L) distance $d(\hat{\boldsymbol{\theta}}_m,\boldsymbol{\theta}_*)$. where $\boldsymbol{\theta}_*$ represents the parameters for the "true" or *generating model*. The Akaike Information Criterion (AIC) [1] approximates $d(\hat{\boldsymbol{\theta}}_m,\boldsymbol{\theta}_*)$ by

$$d(\hat{\boldsymbol{\theta}}_m,\boldsymbol{\theta}_*) \approx -2\log L(\hat{\boldsymbol{\theta}}_m) + 2d_m. \tag{9}$$

While AIC has not been used in surface reconstruction, [9] uses a popular variant of AIC, CAIC [10]

$$d(\hat{\boldsymbol{\theta}}_m,\boldsymbol{\theta}_*) \approx -2\log L(\hat{\boldsymbol{\theta}}_m) + d_m(\log n + 1). \tag{10}$$

We study both CAIC and AIC here. When σ is unknown, $L(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}_m)$ replaces $L(\hat{\boldsymbol{\theta}}_m)$ in (9) and (10).

3.1.3 Model selection using MDL principle

A number of model selection criteria are based on the principle of minimizing the total number of bits to express the observed data. The number of bits required to express the observed data using model m is $\mathrm{len}_m = \mathrm{len}(\hat{\boldsymbol{e}}_m) + \mathrm{len}(\hat{\boldsymbol{\theta}}_m)$, where len denotes the length of the bit string required to encode any quantity. Model selection criteria based on the MDL principle choose the model that minimizes len_m. The quantities $\mathrm{len}(\hat{\boldsymbol{e}}_m)$ and $\mathrm{len}(\hat{\boldsymbol{\theta}}_m)$ are calculated using different assumptions, giving rise to different model selection criteria. The most common of these criteria is due to Rissanen [36], and is equivalent to BIC, eq. (8). In [37], Rissanen derived an improved criterion which chooses the model minimizing

$$\mathrm{len}_m = -\log_2 L(\hat{\boldsymbol{\theta}}_m) + \frac{d_m}{2} \log_{2^*} \left(\hat{\boldsymbol{\theta}}_m^T \, (\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m))^{-1} \, \hat{\boldsymbol{\theta}}_m \right) + \log_{2^*} V_{d_m}, \tag{11}$$

where $\log_{2^*}(t) = \log_2 t + \log_2 \log_2 t + \ldots$, including only its positive terms, and V_{d_m} is the volume of the d_m-dimensional unit hypersphere [18, page 24]. When σ is not known, (11) becomes

$$\mathrm{len}_m = -\log_2 L(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}) + \frac{d_m}{2} \log_{2^*} \left([\hat{\boldsymbol{\theta}}_m^T \, \hat{\sigma}] (\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}))^{-1} \, [\hat{\boldsymbol{\theta}}_m^T \, \hat{\sigma}]^T \right) + \log_{2^*} V_{d_m}. \tag{12}$$

Note again how the measures of accuracy and stability are combined in criteria (11) and (12). In surface reconstruction, a MDL criteria has been used to prune redundant surface by minimizing a quadratic optimization function [29]. Interestingly, this criteria can be shown to be similar in form to AIC [12] which is based on minimizing the K-L distance.

3.1.4 Robust model selection

Information theoretic criteria presented in the last three sections have been traditionally used for data without outliers. This section discusses modifications to the above criteria when outliers are present in the data.

Although the different criteria start from different premises, interestingly, they all end up as a penalized likelihood of the form

$$\log L(\hat{\boldsymbol{\theta}}_m) \ + \ \text{stability or complexity term.} \tag{13}$$

To make such criteria robust in the presence of outliers, the accuracy term or the stability term or both have been modified by different researchers [9, 20, 31, 35, 38, 49, 50]. However, all these modifications are of an empirical nature. A discussion and comparison of these approaches is beyond the scope of the current paper. As such, we only discuss Boyer, Mirza, and Ganguly's [9] modification to CAIC in a surface reconstruction algorithm. Boyer, Mirza, and Ganguly [9] model range data contaminated with outliers to be t-distributed and replace the loglikelihood, $\log L(\boldsymbol{\theta}_m, \sigma)$ in (13), with a weighted loglikelihood function given by

$$\log L(\boldsymbol{\theta}_m, \sigma) \ \propto \ -\sum_{i=1}^{n} w(u_{mi}) \ \rho(u_{mi}) \tag{14}$$

in CAIC, where $\rho(u_{mi})$ is given by

$$\rho(u) = (1 + f) \log{(1 + \frac{u^2}{f})} \quad \text{and} \quad w(u) = \begin{cases} 1 & u = 0 \\ \frac{\psi(u)}{u} & \text{otherwise.} \end{cases}$$

We use this definition of $\log L(\boldsymbol{\theta}_m, \sigma)$ for all the information theoretic criteria whenever data is contaminated with both noise and outliers.

3.2 Model selection using hypothesis tests

A number of model selection criteria that have been used in reconstruction algorithms are based on hypothesis tests. This section summarizes four such criteria. Each starts with the zeroth order model as the null hypotheses and moves to the next higher order model when a null hypotheses is rejected. In these techniques, since all null hypotheses may be rejected, it is possible that no model is selected. This section also introduces a simple, new F-test model selection criteria. This technique may be used when σ is unknown and the Chi-square based techniques cannot be used.

RUNS: The intuition behind using a runs test is that low order incorrect models will produce a large "run" (consecutive sequence) of all positive or all negative residuals. For 2D range images, the total number of runs, r_m, for any fit $\hat{\boldsymbol{\theta}}_m$, is asymptotically[2] normally distributed and is given by [11, pages 164-170]

$$r_m \sim N\left(\frac{2p_m q_m}{p_m + q_m} + 1, \quad \frac{2p_m q_m(2p_m q_m - p_m - q_m)}{(p_m + q_m)^2(p_m + q_m - 1)}\right).$$

Here, p_m is the number of positive residuals, and n_m is the number of negative residuals in the fit. The test rejects model m if r_m is not within a 95% level of confidence. Since the RUNS test does not generalize to 3D range images, Besl [4, pages 150-152] introduces a heuristic approximation. He creates binary images of positive and negative residuals, erodes the images using a 3×3 kernel, finds the largest connected component in each image, and rejects the null hypotheses if the larger of these components is greater than 2% of n. We follow this heuristic for 3D range images. The runs test is advantageous when σ and the noise distribution of the data are unknown.

CHI: This test is based on a one-way Chi-square test and rejects model m at a 95% confidence level. It has been used by Whaite and Ferrie [52]. The intuition is that low order incorrect models

[2] For small samples, techniques from [46] and [21] may be used.

will produce a significant over-estimate to the error in the data.

CR-Test: This test combines CHI and RUNS, and rejects model m if both of them fail. This is the model selection criteria used by Besl and Jain [6].

CSR-Test: This test, based on Bolles and Fischler's [8] test, rejects model m for any one of three reasons: (a) CHI fails, (b) Reject m at 95% confidence level when $|p_m - n_m| > 2\sqrt{n}$, and (c) Reject m at 95% confidence level when the longest run exceeds $3.32 + \log_2 n$. For 3D range images, we replace the longest run with size of the largest connected component created by the process described in RUNS.

FTEST: In this test, any model m_i is rejected in favor of m_{i+1} if [51, page96]

$$\frac{(RSS_{m_i} - RSS_{m_{i+1}})/((n-d_{m_i}) - (n-d_{m_{i+1}}))}{RSS_{m_i}/(n-d_{m_{i+1}})} > F_{(d_{m_{i+1}} - d_{m_i}, n-d_{m_{i+1}});0.95}. \tag{15}$$

Starting with the zeroth-order model, this test continues switching to a higher order model until (15) is not satisfied or until all models in M have been tested.

4 Model selection using bootstrap principle

The model selection criteria presented in sec. 3, with the exception of RUNS, implicitly assume that the error distribution is known *a priori*. Some, such as the techniques based on chi-square tests and F-test in sec. 3.2 are more restrictive, specifically assuming a Gaussian distribution. In computer vision problems, however, error distributions are often unknown and difficult to model accurately, making it crucial to develop model selection criteria that depend on only weak assumptions about error distributions. We address this problem in this section by deriving bootstrap [19] versions of model selection criteria in sec. 3.1. The resulting criteria are empirical in nature, making them somewhat expensive to compute, but they do not require user-defined thresholds and can be used when sensor error models are unavailable or unreliable.

The bootstrap is a method for estimating an unknown distribution from available data. This technique introduced in statistics by Efron [19], has only recently been used in computer vision [14]. In regression, the bootstrap technique can be used to obtain an empirical distribution of errors in the data, and this distribution can be used to generate different statistics on the measured depth values, $\boldsymbol{z}$, and the estimated parameter vector, $\boldsymbol{\theta}_m$. As discussed later, we will need bootstrap estimates of the standard deviation of noise, σ, and the covariance matrix, $\boldsymbol{V}(\boldsymbol{\theta}_m)$ for the different model selection criteria. The idea of bootstrap is simple. Consider the regression model given by (3). Let $\hat{\boldsymbol{\theta}}_m$ be the parameter estimate of $\boldsymbol{\theta}_m$, let $\hat{\boldsymbol{e}}_m$ be the corresponding residuals, and let $\hat{\boldsymbol{z}}_m$ be the vector of estimated z values. The residuals, $\hat{\boldsymbol{e}}_m = [\hat{e}_{m1} \ldots \hat{e}_{mn}]^T$, can be used to generate $\hat{P}_m$, an empirical distribution function. The plug-in bootstrap principle [19, chapter 4] samples from $\hat{P}_m$ to generate bootstrap data. Note that sampling from $\hat{P}_m$ is the same as sampling from the set $\{\hat{e}_{m1}, \ldots, \hat{e}_{mn}\}$ with replacement. The bootstrap error vector $\boldsymbol{e}^*_m$ is added to $\hat{\boldsymbol{z}}_m$ to generate a bootstrap set of z values, $\boldsymbol{z}^*_m$. This "bootstrap data" set can now be used to generate a bootstrap estimate of $\hat{\boldsymbol{\theta}}^*_m$. The process is repeated to generate R bootstrap error vectors $\boldsymbol{e}^*_{1m}, \ldots, \boldsymbol{e}^*_{Rm}$, adding each $\boldsymbol{e}^*_{km}$ to $\hat{\boldsymbol{z}}_m$ gives R bootstrap response vectors $\boldsymbol{z}^*_{1m}, \ldots, \boldsymbol{z}^*_{Rm}$. These in turn can be used to generate bootstrap estimates $\hat{\boldsymbol{\theta}}^*_{1m}, \ldots, \hat{\boldsymbol{\theta}}^*_{Rm}$. The response vectors $\boldsymbol{z}^*_{1m}, \ldots, \boldsymbol{z}^*_{Rm}$ are used to generate a bootstrap estimate of σ, and $\hat{\boldsymbol{\theta}}^*_{1m}, \ldots, \hat{\boldsymbol{\theta}}^*_{Rm}$ are used to generate a bootstrap estimate of $\boldsymbol{V}(\boldsymbol{\theta}_m)$. Figure 4 gives a schematic diagram of the bootstrap technique.

Note that the above description does not specify a method for estimating $\hat{\boldsymbol{\theta}}_m$. The bootstrap method is general and may be used for data contaminated with noise and outliers. It is based on the assumption that errors are independent, and the lack of independence generally reduces the accuracy of the result [19, page 396]. The number of bootstrap replications, R, is chosen empirically. According to Efron and Tibshirani [19, page 52], seldom are more than 200 replications needed for estimating the mean and covariance.

4.1 Behavior of bootstrap estimates of spread

This section only uses bootstrap measures of spread to derive bootstrap versions of model selection criteria in sec. 3.1. In particular, the section uses bootstrap estimate of σ and the bootstrap estimate of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$ in several information theoretic model selection criteria.

The bootstrap estimate of σ, σ^*_m, is calculated by finding the average standard deviation of $\boldsymbol{z}^*_{1m}$, $\ldots$, $\boldsymbol{z}^*_{Rm}$, and the bootstrap estimate of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$ is given by the covariance matrix of $\hat{\boldsymbol{\theta}}^*_{1m}, \ldots, \hat{\boldsymbol{\theta}}^*_{Rm}$. To study the behavior of the bootstrap estimates of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$, data from Gaussian distribution is used, and $\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)$ corresponding to each model is compared with the expected covariance matrix $E(\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)) = \sigma^2 \left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1}$.

The first set of experiments generate data from the models $z = 100 + x$ and $z = 100 + x - 0.1x^2$ at $\sigma = 0.05$. Each experiment increases the region size symmetrically around the origin from 7 to 77 pixels. Figure 5(a) and (b) show the bootstrap estimates, $\sigma^*_{m_0}$, $\sigma^*_{m_1}$, $\sigma^*_{m_2}$, and $\sigma^*_{m_3}$. The results show that none of the σ^*_m values are close to the actual σ at small region sizes. However, as region size increases, $\sigma^*_{m_1}$, $\sigma^*_{m_2}$, and $\sigma^*_{m_3}$ gradually approach the actual σ value in fig. 5(a), giving a reasonably accurate estimate of σ beyond a region size of 30 pixels. Similarly, in fig. 5(b), σ^*_{m2}, and

$\sigma^*_{m_3}$ give a reasonably accurate estimate of σ beyond a region size of 30 pixels. But in both cases, σ^*_m values for models of lower order than the correct model are gross overestimates of σ. Figures 6 (a) and (b) compare the corresponding bootstrap estimates of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$ by plotting $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ against $\log|\sigma^2\left(\boldsymbol{X}_m^T\boldsymbol{X}_m\right)^{-1}|$ for the zeroth, linear, quadratic, and cubic models. The results show that $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ is close to $\log|\sigma^2\left(\boldsymbol{X}_m^T\boldsymbol{X}_m\right)^{-1}|$ for the correct model and models of higher order than the correct model. For models of lower order than the correct model, $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ is overestimated. This is expected because bootstrap estimates of σ for lower order models are overestimated, and consequently, $|\sigma^2\left(\boldsymbol{X}_m^T\boldsymbol{X}_m\right)^{-1}|$ is overestimated. (Recall that a higher value of $|\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)|$ implies a more stable model). This implies that bootstrap estimates of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$ for lower order models may bias a model selection criteria towards lower order models. As shown later, this bias is effectively compensated by the accuracy term and does not pose a problem [12].

4.2 Bootstrap criteria for data without outliers

As mentioned in sec. 3.1.4, the information theoretic criteria are in the form of a penalized likelihood, balancing between accuracy of the fit and stability or complexity of the model. When error distributions are unknown, the two quantities must be approximated using bootstrap statistics and based on weak assumptions regarding the data. First consider the accuracy term given by the likelihood. When errors are unknown and do not contain outliers, OLS is used for parameter estimation. More sophisticated estimators, such as MLE or M-estimators cannot be used because they assume specific error distributions. Besides, these estimators are not necessary because OLS gives unbiased, minimum variance estimates [32, page 172] even with our weak assumptions about sensor noise. The accuracy of the model given the data may then be measured using the normalized residual sum of squares, making prior knowledge of error distributions unnecessary. Therefore, we replace the model accuracy term $\log L(\hat{\boldsymbol{\theta}}_m)$ with $-RSS_m/\sigma^2$. But, σ is still unknown. As demonstrated in sec. 4.1, however, σ^*_m values estimated using the correct model and those using any model of higher order than the correct model are close to each other and to the true σ. Thus, for $M = \{m_0,\ m_1,\ m_2,\ m_3\}$, $\sigma^*_{m_3}$ can be used as a good estimate of σ. As such, the accuracy term in the bootstrap criteria is given by $-RSS_m/{\sigma^*_{m_3}}^2$.

The stability or complexity term for the information theoretic criteria requires only d_m for AIC, d_m and n for BIC and CAIC, making each independent of the error distribution. For bootstrap versions of BAYES and RISS the stability or complexity term depends on $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$ which is replaced by its bootstrap estimate $\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)$. The bootstrapped Bayesian model selection criterion, which we call, BMSC-BAYES, is obtained by taking the natural logarithm of (5), replacing $\log L(\hat{\boldsymbol{\theta}}_m)$ with $-RSS_m/{\sigma^*_{m_3}}^2$ and $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_m)$ with $\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)$, yielding

$$\text{BMSC-BAYES}_m = \frac{d_m}{2}\log 2\pi \ - \ \frac{RSS_m}{{\sigma^*_{m_3}}^2} \ + \ \frac{1}{2}\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|. \tag{16}$$

Similarly, RISS (11) can be approximated as

$$\text{BMSC-RISS}_m = \frac{RSS_m}{{\sigma^*_{m_3}}^2} \ + \ \frac{d_m}{2}\log_2 * \left(\hat{\boldsymbol{\theta}}_m^T \ |\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|^{-1} \ \hat{\boldsymbol{\theta}}_m\right) \ + \ log_2 * (V_{d_m}). \tag{17}$$

Note that while BMSC-BAYES needs to be maximized, BMSC-RISS needs to be minimized. These criteria can be used when error characteristics are unknown or unreliable.

4.3 Bootstrap criteria for data with outliers

When data is contaminated by outliers, $\boldsymbol{\theta}_m$ cannot be estimated using OLS, and robust parameter estimation techniques must be used. Although the bootstrap principle is independent of parameter estimation techniques, most robust estimation techniques, such as M-estimators, make strong assumptions regarding the noise in the data. As such, these techniques cannot be used when error distributions are unreliable or unknown. However, LMS (see sec. 1) is robust in the presence of outliers, and only assumes errors are independent and identically distributed. Therefore, LMS is used for estimating $\hat{\boldsymbol{\theta}}_m$ and the bootstrap parameters $\hat{\boldsymbol{\theta}}^*_{1m},\ \ldots,\ \hat{\boldsymbol{\theta}}^*_{Rm}$. This is not sufficient, however: both the accuracy and stability terms behave differently in the presence of outliers, so neither (16) nor (17) may be used for model selection.

Let us consider, in turn, the measures of accuracy and stability. First, consider the accuracy term given by $RSS_m/{\sigma^*_{m_3}}^2$. The denominator term, $\sigma^*_{m_3}$, cannot be accurately estimated in the presence of outliers. Recall from sec. 4.1 that σ^*_m is calculated by finding the average standard deviation of $\boldsymbol{z}^*_{1m}, \ldots, \boldsymbol{z}^*_{Rm}$. However, this approach cannot be used in the presence of outliers because some z^*s in $\boldsymbol{z}^*_{jm}$ will correspond to outliers while others will correspond to inliers. Calculating σ^*_m by finding the average standard deviation using z^*s in $\boldsymbol{z}^*_{jm}$ which are inliers does not give a reliable estimate of σ, yielding worse estimates with increase in the percentage of outliers. In this section, therefore, we simply use RSS_m normalized by the number of degrees of freedom, $(n - d_m)$, as the accuracy term in the model selection criteria.

For the stability term, we have seen in sec. 4.2 that in the absence of outliers $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ is close to $\log|\sigma^2\left(\boldsymbol{X}_m^T\boldsymbol{X}_m\right)^{-1}|$ for the correct model and for models of higher order than the correct model. For models of lower order than the correct model, $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ is overestimated. However, as the fraction of outliers in the data increases, the correct model and higher order models start overestimating $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ [12]. But, this moderate increase in $\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|$ does not bias the criteria towards any particular model. As such, measures of stability used in the bootstrap criteria of sec. 4.2 are left unaltered.

Thus, when the data contains outliers, the bootstrap criteria are obtained by replacing RSS_m/σ^*_{m3} with $RSS_m/(n-d_m)$ in (16) and (17):

$$\text{BMSC-BAYES}_m = \frac{d_m}{2}\log 2\pi \; - \; \frac{RSS_m}{n-d_m} \; + \; \frac{1}{2}\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|, \quad \text{and} \tag{18}$$

$$\text{BMSC-RISS}_m = \frac{RSS_m}{n-d_m} \; + \; \frac{d_m}{2}\log_2 * \left(\hat{\boldsymbol{\theta}}_m^T \, |\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|^{-1} \, \hat{\boldsymbol{\theta}}_m\right) \; + \; log_2 * (V_{d_m}). \tag{19}$$

Figures 7 shows how BMSC-BAYES balances between accuracy and stability. Figures 7(a) and (b) show the interaction between accuracy and stability terms with about 5% outliers in the data. Observe the large jumps in the accuracy term until the correct model is reached, after which the accuracy term does not change much. It is left to the stability term to distinguish between the correct model and other higher order models. This behavior is repeated in figs. 7(c) and (d) with 30% outliers.

5 New rules for surface merging

This section extends the model selection framework to develop new rules for merging surface patches to a single surface description. We assume that small surface patches have already been estimated using different approaches summarized in the introduction, and these surface patches do not undersegment the scene, i.e, they do not bridge discontinuities. For the following discussion, we only consider the core problem of merging two surface patches.

To define the problem precisely, suppose one surface, A, is fit to data set D_A and another, B, is fit to data set D_B where $D_A \bigcap D_B = \emptyset$. The issue is to determine whether D_A and D_B are measurements from the same or different underlying surfaces. When D_A and D_B are measurements from the same surface, they should be merged into a single surface, C, which can use any model $m \in M$. Let $C_0, \ldots, C_3$, corresponding to models $m_0, \ldots, m_3$, be fits to the data set $D_C = D_A \bigcup D_B$. Surface merging involves a choice between selecting $\{A, B\}$ for D_A and D_B or any one of C_0, C_1, C_2, and C_3 for D_C.

5.1 New rules based on information-theoretic criteria

As seen in sec. 3.1, different information theoretic criteria compare Bayesian probabilities, K-L distances, or minimum description lengths to select the best model from m_0, m_1, m_2, and m_3. For surface merging, we extend this notion, and compare the same quantities to select models m_A and m_B together, which preserves the two separate surfaces, or to select any one of models m_0, $\ldots$, m_3 for the data set $D_C = D_A U D_B$, thereby, merging the two surfaces.To do this, measures of probabilities, K-L distances, and description lengths must be formulated for m_A and m_B combined. Since D_A and D_B are disjoint, $P(D_A D_B|m_A, m_B, I) = P(D_A|m_A, I)\ P(D_B|m_B, I)$. Similarly, evaluating at maximum likelihood estimates the K-L distance also reduces to $d(\hat{\boldsymbol{\theta}}_A, \boldsymbol{\theta}_*) + d(\hat{\boldsymbol{\theta}}_B, \boldsymbol{\theta}_*)$. Finally, in the MDL case len_{m_A, m_B} is simply equal to $len_{m_A} + len_{m_B}$. Based on this, merging decisions for Bayesian probabilities, K-L distances, and MDLs may be represented as

$$\max\{(P(D_A|m_A, I)P(D_B|m_B, I)), P(D|m_{m_0}, I), \ldots, P(D|m_{m_3}, I)\},$$

$$\min\{(d(\hat{\boldsymbol{\theta}}_A, \boldsymbol{\theta}_*) + d(\hat{\boldsymbol{\theta}}_B, \boldsymbol{\theta}_*)), d(\hat{\boldsymbol{\theta}}_{m_0}, \boldsymbol{\theta}_*), \ldots, d(\hat{\boldsymbol{\theta}}_{m_3}, \boldsymbol{\theta}_*)\},$$

$$\min\{(len_{m_A} + len_{m_B}),\ len_{m_0},\ \ldots, len_{m_3}\},$$

respectively. Replacing the model selection criteria of sec. 3 in the appropriate decision functions above, we get merging rules based on AIC[3], CAIC, BAYES[4], BIC, and RISS. For formulating merging rule based on bootstrap criteria BMSC-BAYES and BMSC-RISS, note that they are based on the logarithm of the Bayesian probability, $P(D|m, I)$, and MDL, len_m, respectively. As such, the corresponding merging rules are similar to the Bayesian and MDL rules:

$$\max\{(\text{BMSC-BAYES}_A + \text{BMSC-BAYES}_B), \text{BMSC-BAYES}_{m_0}, \ldots, \text{BMSC-BAYES}_{m_3}\}$$
$$\max\{(\text{BMSC-RISS}_A + \text{BMSC-RISS}_B), \text{BMSC-RISS}_{m_0}, \ldots, \text{BMSC-RISS}_{m_3}\}$$

5.2 Merging rules from hypothesis tests

This section formulates simple merging rules using the hypothesis testing criteria discussed in sec. 3.2. The tests RUNS, CHI, CR, and CSR may each reject all candidate models, and therefore, it is possible that no model is selected. Based on this each may be extended to a merging rule that merges A and B to C if and only if a model from M is selected for C. Note how these rules do not use any information from the fitted surfaces A and B.

Our merging rule based on the F-test works in two steps. In the first step, it checks if the parameters of surface A are within the 95% confidence interval of the parameters of B (or vice-

[3] It can be shown that the optimization function used in [29] may be used for merging surfaces, and is similar to the merging rule based on AIC [12].

[4] In surface reconstruction, a similar Bayesian merging approach has been used in [27]. However, this approach only merges surfaces corresponding to the same model. Besides, it also constrains the parameter space so that $||\boldsymbol{\theta}_m|| = 1$. As such, this work can be considered as a special case of ours.

versa — only one must succeed) [44] using the F statistic in [51, page 97]. When A and B belong to different models, the technique only checks if the lower order model fits within the confidence interval of the higher order model. If this step decides that the surfaces be merged, then the second step uses FTEST of sec. 3.2 to find the best model .

6 Factors affecting performance

A large number of model selection criteria were presented in sections 3 and 4. These criteria are categorized in fig. 8. Sec. 5 formulated new merging rules based on these criteria. The next step is to analyze this wide variety of model selection criteria and merging rules. This analysis must study the effects of several different influences on the performance of model selection criteria and merging rules.

1. **Region size**: It is easier to identify the correct model over a relatively larger region size. Figures 9(a) and (b) show data from a quadratic model at region sizes (a) 25 pixels, and (b) 50 pixels. Observe, how the data points in fig. 9(a) appear to be from a line.

2. **Underlying surface**: It is difficult to identify the correct model for surfaces having small magnitude parameters for the highest order term. Figure 10 shows noisy data points from the line $y = 100 + x + a_2x^2$, at (a) $a_2 = -0.02$, (b) $a_2 = -0.1$, (c) $a_2 = -0.5$. Observe how data in figs. 10(a) and (b) appear to be from a linear model.

3. **Noise level**: It becomes difficult to select the correct model with increasing noise in data. Figures 11(a) and (b) illustrate this point with data from a quadratic model with Gaussian

noise at $\sigma = 0.02$ and $\sigma = 0.1$, respectively. While it is easy to identify the quadratic model at $\sigma = 0.02$ (fig. 11(a)), the data points in fig. 11(b) appear to be from a linear model. Similarly, it is harder to detect a discontinuity with increase noise in the data (see fig. 12).

4. **Number of alternative models**: Model selection criteria may be biased towards lower (or higher) order fits. Such a bias may not be detected, for example, with data from a quadratic fit (model m_2), if M only consists of quadratic and cubic models, m_2 and m_3. To detect such biases, $M = \{m_0,\ m_1,\ m_2,\ m_3\}$, is used for experiments with data from linear and quadratic fits.

   The same problem occurs with merging rules. Figure 2(a) shows the correct representation for the data points. However, fig. 2(c) also appears to be a good representation for the data. In this situation, a merging rule may incorrectly choose to merge the surfaces to a single quadratic surface. However, if an application only fits lines to the data, the choice is only between figs. 2(a) and (b), and the discontinuity is likely to be preserved. The experiments demonstrate this by using different sets of candidate models.

5. **Type and magnitude of discontinuities**: A good merging criteria must detect small magnitude discontinuities, and also correctly merge artificial (non-existent) discontinuities. The performance of merging criteria is characterized by testing them over a wide range of step and crease discontinuities, as well as artificial discontinuities.

7 Experimental analysis

Based on the above discussion, a wide range of experiments were conducted to characterize the performance of all criteria over all the factors influencing performance. Experiments are conducted on both synthetic and sensor data. While synthetic data allows us to test different criteria over a exhaustive range of experimental conditions, sensor data allows us to test performance with data containing sensor noise, outliers, and potentially, other kinds of unmodeled errors. This section discusses some of these experiments, and summarizes the performance of various model selection criteria and merging rules. For details of this experimental analysis refer [12].

7.1 Simulation Results

The model selection criteria in fig. 8 and the merging rules based on them, make different assumptions about the data. In order to study the relative performance of all criteria under a common data set, the simulations use a Gaussian noise model and provide noise variance to the criteria that need it. The data points are generated using a perspective projection model with focal length=1.77 cm and pixel size = 16μ m, the calibration parameters of our range sensor [40]. The results are based on 500 simulations, and for bootstrap criteria, the number of bootstrap replications, R, is set to 200 (see sec. 4). Unless mentioned otherwise, $M = \{m_0,\ m_1,\ m_2,\ m_3\}$.

7.1.1 Model selection

The experiments are based on data sets from linear and quadratic models given by $z = a_0 + a_1x$, and $z = a_0 + a_1x + a_2x^2$, respectively.

Effect of region size and σ on performance: In the first set of experiments, $a_0 = 100$ and $a_1 = 1$ (for both the models), and $a_2 = -0.1$ for the quadratic model. The region size is increased from 7 pixels to 77 pixels, symmetrically around the origin, and σ is varied from 0.02 cm to 0.1 cm. Figure 13 shows percentage success of different selection criteria for data from the linear model at $\sigma = 0.05$ cm. Figure 13(a) shows results for BAYES, RISS, AIC, BIC, CAIC, and BMSC-BAYES, and fig. 13(b) shows results for RUNS, CHI, CSR-test, CR-test, F-test, and BMSC-RISS. The results in fig. 13(a) show that RISS performs the best, and although BAYES, BIC, and CAIC work poorly at small region sizes, their performance improves as region size increases. The new bootstrap based criteria, BMSC-BAYES also performs well, closely following BAYES. This performance is promising, given that it does not make any assumption regarding the noise distribution. The criteria based on hypothesis tests have a success rate from 90 to 95%. This is expected because they are based on a 95% confidence interval. Surprisingly, however, AIC shows a success rate of only 80%, and tends to choose quadratic and cubic fits over a linear fit. Although not shown here, the results exhibit small improvements at $\sigma = 0.02$ and small degradations at $\sigma = 0.1$.

Figure 14 shows corresponding performance for data from the quadratic model. The figure show that the results are poor for all criteria at small region sizes, and show close to "steady state" performance (say, within 3% of maximum success rate) after a certain *minimum region size*. This minimum region size changes with σ. For example, for BAYES this size is 25 pixels at $\sigma = 0.02$ cm and 40 pixels at $\sigma = 0.1$ cm. The results show several differences from the linear case. First, with increasing σ, all criteria show worse performance for the quadratic model at small region sizes. This is not surprising given the difficulty of seeing a quadratic fit in the data in Figure 11(b). Second, RISS and BMSC-RISS, which perform the best for linear models even for small regions, now perform poorly at small region sizes. This suggests that RISS and BMSC-RISS are biased towards low order surfaces. But once again, at large region sizes BAYES, RISS, and BAYES-BMSC perform the best. CAIC and BIC closely follow these criteria. AIC again shows a success rate of about 80%, and tends toward choosing cubic fits.

Effect of changing a_1 and a_2: In this set of experiments, we vary a_1 for the linear model, (keeping a_0 fixed), and vary a_2 for the quadratic model (keeping a_0 and a_1 fixed), at $\sigma = 0.05$ cm and a region size of 25 pixels. All criteria show poor results at small magnitudes of a_1 and a_2 (refer Figure 10). The relative performance of different criteria remains the same as above large magnitudes of a_1 and a_2 [12].

Overall performance: To summarize, most criteria perform well at moderate region sizes (greater than 25 pixels) under moderate noise levels, and the performance of all criteria drops at small region sizes, high values of σ, and at low magnitudes of a_1 and a_2. Intuitively, the results match our own ability to detect models from sample data in figs. 9, 10, and 11. As far as specific criteria are concerned, BAYES, BMSC-BAYES, and CAIC perform the best, the performance of BIC is only slightly worse. BAYES and BMSC-BAYES outperform CAIC at larger region sizes. AIC seems to overfit, while RISS shows a slight bias towards lower order surfaces. The second column of table 3 gives a qualitative summary of relative performance.

7.1.2 Surface Merging

This section compares the performance of different merging rules introduced in sec. 5 on surface fits with step and crease discontinuities (see Figure 15), and artificial discontinuities (formed when $h = 0$ or $\alpha = 0$). The experiments are based on data generated from linear models.

Step discontinuities: For step discontinuities, data are generated from the following two surfaces:

$$A : z = (100 - \frac{h}{2}) + x, \quad B : z = (100 + \frac{h}{2}) + x.$$

Thus, A and B are separated by a step height of h cm.

Performance at different step heights at a relatively small region size: It is difficult to preserve small magnitude discontinuities at small region sizes. However, some merging rules may perform better than others when the region size is small. Figure 16 shows the percentage success of merging rules in detecting a discontinuity at different values of h/σ when each surface has a region size of 25 pixels. Figure 16(a) shows the performance of merging rules based on BAYES, RISS, AIC, BIC, CAIC, and BMSC-BAYES, while fig. 16(b) shows the performance of merging rules based on RUNS, CHI, CSR-test, CR-test, F-test, and BMSC-RISS. The results show that merging rules based on AIC, BIC, CAIC, BAYES, BMSC-BAYES, and F-test perform extremely well, even at such a small region size. These rules detect discontinuities with 98% success at $h = 3\sigma$ and 100% success at $h \geq 4\sigma$. In contrast to these criteria, RISS, CHI, CSR-test, CR-test, and BMSC-RISS perform relatively poorly. RISS, CHI, CSR-test, and CR-test require $h = 6\sigma$ for 100% success, BMSC-RISS requires $h = 8\sigma$, and RUNS requires $h = 12\sigma$. Thus, AIC, BIC, CAIC, BAYES, BMSC-BAYES, F-test clearly show better performance than other merging rules. Observe how the merging rule based on the newly introduced BMSC-BAYES performs well, and is only slightly worse than the merging rule based on BAYES.

Performance with increasing region sizes at small magnitude step heights: The above set of experiments showed that at a region size of about 25 pixels and $\sigma = 005$ cm, even the best merging rules perform well only when the step height is greater then $h = 3\sigma$. This set of experiments study the performance of merging criteria at $h = 2\sigma$ when the region size is increased from 36 pixels to 102 pixels. The results (fig. 17(a) and (b)) show that AIC, BIC, CAIC, BAYES, and BMSC-BAYES detect a discontinuity with 100% success at a region size of 85 pixels. Thus, given a sufficiently large region size, these criteria can even detect such a small magnitude discontinuity. RISS and BMSC-RISS show poor performance, detecting the discontinuity with 44.2% and 8% success even at a region size of 102 pixels. This suggests that these criteria are slightly biased towards merging surfaces. Likewise, RUNS, CHI, CR-test, and CSR-test also show only 26.2%, 67.6%, 75.6%, and 85.8% success at 102 pixels. Surprisingly, F-test shows almost 100% success beyond a region size of 40 pixels, suggesting a possible bias, confirmed later, toward preserving discontinuities.

Crease discontinuities: For crease discontinuities, we generate data from the following two equations (fig. 15(b)):

$$A : z = 100 + x \tan\left(\frac{\pi}{4} + \alpha\right), \quad B : z = 100 + x \tan\left(\frac{\pi}{4} - \alpha\right).$$

Performance at different crease angles at a relatively small region size: As in the case of step discontinuities, fig. 18 shows percentage success of merging rules in detecting a discontinuity at different values of α when each surface has a region size of 25 pixels and $\sigma = 0.05$. The results show the same performance trends as for the step discontinuity. Merging rules based on AIC, BIC, CAIC, BAYES, BMSC-BAYES, and F-test perform well, even at such a small region size. The merging rule based on F-test shows a 100% success at $\alpha = 4$ degrees, while merging rules based on AIC, BIC, CAIC, BAYES, BMSC-BAYES show 98% success at $\alpha = 6$ degrees and a 100% success at $\alpha = 8$ degrees. Among other merging rules, CHI, CSR-test, and CR-test show a 100% success at $\alpha = 10$ degrees, RISS at $\alpha = 11$ degrees, BMSC-RISS at $\alpha = 12$ degrees, and RUNS at $\alpha = 15$ degrees. Table 1 shows these α values at different σ.

Performance at different region sizes at small magnitude crease angles: The above set of experiments showed that at a region size of about 25 pixels, even the best merging rules perform well only when α is greater than 6 degrees. This set of experiments study the performance of merging rules at extremely small magnitude crease discontinuities when the region size for each surface is increased from 36 pixels to 102 pixels. Figures 19(a) and (b) show the performance of merging rules with increasing region size at $\alpha = 2$ degree. The results show that AIC, BIC, CAIC, BAYES, and BMSC-BAYES detect a discontinuity with 100% success at a region size of about 60 pixels. Note how all these rules can detect such a small magnitude discontinuity, when given data from a sufficiently large region size. On the other hand, F-test always shows a 100% success, while RISS and BMSC-RISS show 100% success beyond region sizes of 78 and 84 pixels, respectively. CHI and CR-test show 100% success beyond region sizes of 90 pixels, while CSR-test shows a 100% success at a region size of 78 pixels. RUNS performs the worst, showing 98.2% success even at a region size of 102 pixels.

Performance with changing model set: Figure 2 illustrated the difficulty in distinguishing between a quadratic model and a low magnitude crease discontinuity. To study this point further, the next set of experiments studies the relative performance of merging rules by using $M = \{m_1\}$ at a region size of 25 pixels and $\sigma = 0.05$ cm. Observe the improved performance in the results shown in fig. 18(a) and (b). Here, AIC, BIC, CAIC, BAYES, and BMSC-BAYES show 100% success at $\alpha = 2$ degrees. This is a significant improvement compared to 100% success at $\alpha = 8$ degrees when $M = \{m_0,\ m_1,\ m_2,\ m_3\}$. Among other merging rules, RISS, CHI, CSR-test, and CR-test show a 100% success at about $\alpha = 3$ degrees, and BMSC-RISS and RUNS at $\alpha = 4$

degrees. Thus, when $M = \{m_1\}$, although there is no explicit model selection, the merging rules based on model selection criteria give significantly improved performance for data from crease discontinuities, even at a relatively small region size.

Non-existent discontinuities: For artificial (non-existent) discontinuities, data are generated for surfaces A and B from the line $z = 100 + x$ at a region size of 25 pixels per surface. Table 2 shows the performance of different merging rules. The results show that merging rule based on RISS, BAYES, and BMSC-BAYES perform the best, followed by RUNS, CAIC, CHI. Merging rules based on BIC, CR-test, and CSR-test show a modest success rate of about 91%. A couple of points must be mentioned here. First, RISS and BMSC-RISS which showed poor performance when detecting actual discontinuities, show 100% success when merging artificial discontinuities. This suggests a bias in these criteria towards merging surface fits. Second, BIC, whose performance is comparable to that of BAYES, BMSC-BAYES and CAIC in preserving actual discontinuities, is only a 91% success when merging artificial discontinuities. This suggests BIC is slightly biased towards preserving discontinuities. Among other merging rules, F-test, which performed well when detecting discontinuities, only merges 15.4% of the artificial discontinuities, suggesting it is strongly biased toward preserving discontinuities. AIC shows only a 76.6% success. This is because, although AIC merges artificial discontinuities, it merges them to higher order surfaces. The merging rules show some improvement in results with increasing region size, and showed no change in performance when σ was varied from 0.02 cm to 0.1 cm [12].

Overall performance: To summarize, the performance of BAYES, BMSC-BAYES, and CAIC is the best, and that of BIC is only slightly worse. These criteria work well even at relatively small region sizes (25 pixels). As region size increases, the criteria can detect extremely small step heights ($h = 2\sigma$ at 85 pixels) and crease angles ($\alpha = 2$ degrees at 60 pixels). BMSC-BAYES shows consistently good results and gives a useful merging rule when sensor error models are unavailable or unreliable. Among other criteria, RISS and BMSC-RISS, with a bias towards merging surfaces, only show average behavior. F-test and AIC do not perform well for artificial discontinuities. While F-test is biased towards preserving discontinuities, AIC merges to higher order surfaces. RUNS, CHI, CSR-test, and CR-test are also moderately biased towards merging surfaces at small region sizes. This is expected because these merging rules do not use any information from the old fits, but only look at finding a possible model for the combined data set. Finally, merging rules perform extremely well at crease discontinuities when the quadratic model is not present in M. The third column of table 3 gives a qualitative summary of relative performances.

7.2 Results using sensor data

This section compares performance of model selection criteria and merging rules on sensor data. This allows us to test performance with data containing both small-scale random noise and outliers. This noise may often be difficult to model accurately, and potentially, there might be unmodeled errors in the data. Thus, using sensor data, performance may be tested under more realistic conditions. Certainly all criteria in fig. 8 can be used here. However, since sensor data contains outliers, only robust model selection criteria can be used for model selection. As such, only AIC, BIC, CAIC, BAYES, BMSC-BAYES, BMSC-RISS, and RUNS, and merging rules based on them are compared in this section. We compare performance using Perceptron test data sets from the USF Segmentation Comparison Project [25]. The data sets, consisting of planar surfaces, are particularly useful for model selection experiments because ground-truth segments are provided. For model selection, the experiments take data from each ground-truth segment and determine the best model that describes the data using the different model selection criteria. The experiments are repeated using data from test regions of different sizes within certain segments. To test the performance of merging criteria on real discontinuities, adjacent ground truth segments are tested for merging by each merging criteria. Likewise, to test performance non-existent discontinuities, adjacent regions within certain segments are tested for merging by each merging criteria.

Model selection: In the first set of experiments, model selection criteria are applied to data from ground truth segments in the different images. The results are shown for images 20(a) (Image 1) and 21(a) (Image 2) (the corresponding ground truth segments are shown in figs. 20(b) and 21(b)). Table 4(a) shows the ground truth segments identified as non-planar for data from segments in Image 1 using $M = \{m_0,\ m_1,\ m_2,\ m_3\}$. Figure 4(b) shows the corresponding results when M is reduced to $\{m_1, m_2\}$. For understanding the performance of model selection criteria with increasing region sizes, each table is divided into three parts. The first column shows the small segments ($n < (25 \times 25)$) identified incorrectly by the different criteria, while the second and the third columns show the medium ($n < (50 \times 50)$) and large segments ($n >= (50 \times 50)$) identified incorrectly by different criteria. The results show that all criteria fail to identify the correct model at small region sizes, tending to select the lower order model m_0. When m_0 is removed from M, the performance for most criteria improves considerably (Table 4(b)). At medium to large segment sizes, BAYES perform the best, followed by BMSC-BAYES, BIC, CAIC, and RUNS. Observe how BMSC-BAYES identifies all medium and large segments, except segment 20, correctly. Segment 20, close to being normal to the depth axis, is identified as m_0 by BMSC-BAYES; the planar model is selected when m_0 is removed from M in table 4(b). Table 5(a) and (b) show the results for Image

2. The results show poor performance of all criteria at small region sizes. Among these, RISS is most adversely affected, selecting m_0 over m_1 for nearly all small segments in the image. For medium to large segments, BAYES, BMSC-BAYES, RUNS, CAIC, BIC, and RISS perform well. BMSC-RISS again identifies several segments incorrectly, the correct model being selected once m_0 is removed from M. This again shows BMSC-RISS's strong bias towards lower order surfaces.

Overall, the results show several similarities with results using synthetic data. First, all criteria work poorly in small regions. Second, AIC continues to show bias towards higher order surfaces, and RISS and BMSC-RISS show bias towards lower order surfaces. Third, BMSC-BAYES and BAYES consistently perform the best, followed by CAIC, BIC, and RUNS.

A closer look at the performance of BAYES and BMSC-BAYES is warranted. In Image 2, Segment 16 is identified incorrectly by BAYES. However, this is the same segment as segment 17 in Image 1 which is correctly identified as planar by BAYES. (In [12, page 112] we show similar observations for other segments over a large number of images.) All these incorrectly identified segments have one thing in common: the surfaces corresponding to these segments have normals close to the x or y axis, or in other words, these surfaces extend primarily along the depth direction. As such, it is unlikely that the noise distribution of such surfaces can be modeled as t-distributed in the depth direction. BAYES and other information theoretic criteria, being closely tied to the noise distribution of the data, therefore, make errors at such surfaces. The bootstrap principle, evidently, determines a better distribution of the noise in the data, leading to a correct model selection by BMSC-BAYES for these segments. Thus, the experiments clearly demonstrate the usefulness of bootstrap criteria when noise models are inaccurate or unreliable. However, BMSC-BAYES also identifies some segments incorrectly. Observe that several of these segments (20 in Image 1, and

12 in Image 2) are approximately perpendicular to the depth axis, leading to a zeroth-order model being selection over a planar model. These segments are correctly identified as planar when M is reduced to $\{m_1,\ m_2\}$.

The second set of experiments test the different criteria on data from square regions of progressively increasing sizes, starting from the pixels marked 'x' in segments 19, 24, and 37 in fig. 21(b). Once again, all criteria do a poor job in selecting the correct model when the region size is small, and show improved performance as the region size is increased. Table 8 shows the minimum region size required by each criteria in order to select the correct model. The results show that RUNS works well for relatively small region size. AIC follows next, although it quickly starts selecting quadratic and cubic models as region size increases. BAYES, BIC, and CAIC show average performance, while BMSC-BAYES and RISS require a relatively large region region size for selecting the correct model. BMSC-RISS requires the largest region size for all criteria and cannot select the correct model for segment 19. Observe that, in general, the minimum region size required for segment 19 is the largest because it is almost fronto-parallel (perpendicular to the depth axis), and model selection criteria tend to select m_0 over m_1 even for relatively large region sizes. In contrast, segment 37 is almost along the depth axis, and the model selection criteria detect the correct model, even at small region sizes. Observe how BMSC-BAYES which requires relatively large region sizes in segments 19 and 24, detects the correct model with a region size of 22×22 in segment 37. When M is reduced to $\{m_1,\ m_2\}$, all criteria select the correct model beyond a region size of 10×10.

To summarize, the behavior of model selection criteria on sensor data is similar to the simulation results. All criteria work poorly at small region sizes and perform well as segments get larger.

BMSC-BAYES and BAYES perform the best, with RUNS, CAIC and BIC only slightly worse. AIC, RISS, and BMSC-RISS perform poorly. The performance of BMSC-BAYES, which was introduced here, is especially promising, since it may be used when noise models are unavailable or unreliable. Table 9 summarizes the relative performance of the different criteria.

Merging surfaces: This section compares the performance of merging rules on data from different image regions in images 1 and 2. The first set of experiments test the performance of merging criteria on real discontinuities by attempting to merge each ground truth segment with adjacent segments. Table 6(a) and (b) shows the pairs of segments incorrectly merged by the different criteria for Image 1 using $M = \{m_0,\ m_1,\ m_2,\ m_3\}$ and $M = \{m_1\}$, respectively. To study the behavior of merging rules when merging segments of different sizes, each table is divided into three parts. The first column tabulates all incorrect merges involving small segments, the second column tabulates all incorrect merges between medium segments and between medium and large segments, while the third column tabulates all merges between large segments. In these experiments, the results were identical for AIC, BIC, CAIC, BAYES, and RISS. As such, they are together referred to as "info-th" in the tables. The results show these criteria perform extremely well, preserving all the discontinuities in the image. Observe how the segment pairs in the "nut" (involving segments 21, 22, 23, 24) are also preserved by these criteria. BMSC-BAYES merges the extremely small segments, 21 and 23, to adjacent segments. However, it preserves the crease discontinuity formed by segments 22 and 24 inside the nut. BMSC-BAYES preserves all other discontinuities in the image. Similarly, RUNS merges small segments, 21 and 22 to adjacent segments, but preserves the crease discontinuity between segments 23 and 24. All other discontinuities in the image are preserved by RUNS. BMSC-RISS also has problems at small region sizes, but in addition, it merges several medium sized segments to adjacent segments. The results only improve marginally when M is reduced to $\{m_1\}$. Table 7 shows the performance of merging rules on ground-truth segments from Image 2. Once again the information-theoretic criteria perform the best, only merging extremely small segments. BMSC-BAYES is only slightly worse, incorrectly merging segment pairs (11, 40) and (15, 21). BMSC-RISS performs the worst merging a large number of small segments to adjacent segments. It also merges segment pairs (30, 33), (17, 19), and (24, 25), with only marginal improvement in performance even when M is reduced to $\{m_1\}$. This again shows BMSC-RISS's strong bias towards merging surfaces. RUNS shows average performance, merging some small and large segment pairs. However, all discontinuities between large segments are preserved when M is reduced to $\{m_1\}$.

The next experiment tests the performance of merging rules on artificial discontinuities. The merging rules are tested on adjacent regions of progressively increasing region sizes, starting from pixel marked 'x' in segments 19, 24, and 37 in Image 2. At small region sizes, although all rules merge the two surfaces, they do not merge them to the correct model. Other than one or two exceptions, the regions are correctly merged by the criteria when the combined region size reaches those shown in table 8. Most criteria perform reasonably well. However, BMSC-RISS and RISS shows a bias towards lower order surfaces, while AIC, merges regions to a higher order model as the region size increases.

To summarize, most merging rules work well with moderate to large segment sizes, and have problems at small region sizes. Among these, BAYES, CAIC, and BIC perform the best, closely followed by BMSC-BAYES. RUNS shows average performance merging large segments with relatively small magnitude discontinuities. RISS and AIC do not merge surfaces with artificial discontinuities to the correct model at small region sizes, again showing a bias towards lower and higher order surfaces, respectively. BMSC-RISS does not perform well showing a strong bias towards merging surfaces. Based on the above results, table 9 gives a qualitative performance summary of merging rules.

8 Discussion, Summary, and Conclusion

This paper has studied model selection in the context of range image segmentation algorithms. It has characterized the advantages and limitations of existing criteria, introduced promising criteria from the statistics literature, and developed novel bootstrap based criteria using some of them. The paper has formulated theoretically rigorous and effective rules for merging surfaces by extending model selection techniques. The new and existing criteria were compared over a wide range of underlying surfaces, over different region sizes, with different noise levels, using several sets of alternative models, and using both synthetic and sensor data. The results show that although some model selection criteria and merging rules definitely perform better than others, a moderate region size is crucial to the performance of all techniques. Unfortunately, there is no good way of quantifying small, moderate, and large. As rough indicators, a moderate region size is 25 pixels for the simulated data and (25 x 25) pixels for the Perceptron test data. The result also show that BMSC-BAYES introduced in this paper and BAYES adapted from the statistics literature consistently show good performance. The information theoretic merging rules formulated in this paper perform well even at relatively small step sizes ($h = 2\sigma$) and crease discontinuities ($\alpha = 2$ degrees), and consistently merge artificial discontinuities. Unfortunately, none of the model selection criteria and new merging rules work as well as desired. Based on these results, we make the following recommendations when choosing among them.

- When the noise distribution of the data is known or can be closely approximated, BAYES is a good choice for model selection and surface merging. Looking at the qualitative summaries in tables 3, and 9, BAYES shows good performance in all cases. However, BAYES requires estimating $|\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)|$. Therefore, for time-sensitive applications CAIC is a good alternative.
- When noise distribution is not known or cannot be closely approximated, BMSC-BAYES introduced in this paper is a good choice. Although, this technique is computationally expensive, it is easily parellelizable.
- AIC and RISS should in general be avoided.

These results have several implications in improving existing segmentation algorithms, as well as designing new algorithms.

1. Model selection criteria based on confidence intervals, traditionally used in computer vision algorithms [4, 8, 29, 47, 52], should be avoided and information theoretic model selection criteria, preferably BAYES and CAIC, used instead.
2. Existing merging techniques which are based on heuristics and thresholds must be tuned to specific applications. Such techniques should be replaced with the new merging rules to detect small magnitude discontinuities.
3. Model selection and merging do not work well at small region sizes. Segmentation algorithms should only fit, for example, a linear model to small windows and small seed regions,

and use model selection and merging only on moderate to large region sizes.

criteria	min. α (in degrees)			criteria	min. α (in degrees)		
	σ=0.02	$\sigma = 0.05$	$\sigma = 0.1$		σ=0.02	$\sigma = 0.05$	$\sigma = 0.1$
BAYES	3	8	15	RUNS	9	15	27
RISS	4.5	11	18	CHI	4.5	10	21
AIC	3.5	8	15	CSR-test	4	10	18
BIC	3	8	15	CR-test	4	10	18
CAIC	3.5	8	15	F-test	2	4	10
BMSC-BAYES	3	8	15	BMSC-RISS	6	12	24

Table 1: Performance of merging rules at crease discontinuities with changing $\boldsymbol{\sigma}$. Table shows the minimum $\boldsymbol{\alpha}$ required by merging rules to correctly detect a crease discontinuity with 100% success.

rule	% success	rule	% success
BAYES	99.4	RUNS	96.6
RISS	100.0	CHI	94.4
AIC	76.6	CSR-test	90.8
BIC	91.4	CR-test	91.4
CAIC	96.0	F-test	15.4
BMSC-BAYES	98.8	BMSC-RISS	100

Table 2: Percentage success in merging artificial discontinuities to fit from correct model.

	Model selection	Merging rules
Good	BAYES, BMSC-BAYES, CAIC, BIC	BAYES, BMSC-BAYES, CAIC
Average	RISS, BMSC-RISS, RUNS, CHI, CSR-test, CR-test, F-test	BIC, RISS, BMSC-RISS, RUNS, CHI, CSR-test, CR-test
Poor	AIC	AIC, F-test

Table 3: Overall performance of model selection and merging criteria using data with Gaussian errors.

	segments identified incorrectly		
criteria	small	med.	large
AIC	23, 24	22	16
BIC	23, 26		16
CAIC	23, 26		16
BAYES	23, 26		
RISS	21, 23, 26	13, 20	14
BMSC-BAYES	21, 23	20	
BMSC-RISS	26	20	12
RUNS	21, 23		10

(a) $M = \{m_1,\ m_2,\ m_3,\ m_4\}$

segments identified incorrectly		
small	med.	large
	22	16
		16
		16
21, 23		
21, 23		10

(b) $M = \{m_1,\ m_2\}$

Table 4: Model selection results for Image 1.

	segments identified incorrectly		
criteria	small	med.	large
AIC	13, 31, 34, 35, 36, 40, 41	33	10, 16, 18, 25, 28, 29
BIC	13, 31, 34, 35, 36, 40, 41	33	16, 28, 29
CAIC	13, 31, 32, 34, 35, 36, 40, 41		16, 28, 29
BAYES	13, 31, 32, 34, 35, 40, 41		16, 29
RISS	13, 20, 31, 32, 34, 35, 36 38, 40, 41	21, 33	16, 29
BMSC-BAYES	15, 34, 36, 40, 41	21	10, 12, 25
BMSC-RISS	15, 20, 34, 36, 38, 41	21, 33	10, 12, 19, 25, 30
RUNS	34, 35, 36, 38, 40, 41		18, 19, 24

(a) $M = \{m_1,\ m_2,\ m_3,\ m_4\}$

segments identified incorrectly		
small	med.	large
40, 41	33	10, 16, 18, 28
41	33	16, 28
		16, 28
41		16
		16
15, 40, 41		
40, 41		18, 19, 24

(b) $M = \{m_1,\ m_2\}$

Table 5: Model selection results for Image 2.

	segments merged incorrectly		
	merges involving		
criteria	small	med.	large
info-th			
BMSC-BAYES	(17, 21), (17, 23), (21, 22)		
BMSC-RISS	(17, 21), (17, 23), (17,24), (21, 22) (22, 24)	(17, 18) (17, 19) (17, 22)	
RUNS	(17, 21), (21, 22), (22, 24)		

(a) $M = \{m_0,\ m_1,\ m_2,\ m_3\}$

segments merged incorrectly		
merges involving		
small	med.	large
(17, 21), (17, 23)		
(17, 21), (17, 23), (17, 24), (22, 24)	(17, 18), (17, 19) (17, 22)	
(17, 21)		

(b) $M = \{m_1\}$

Table 6: Merging results for Image 1

	segments merged incorrectly		
	merges involving		
criteria	small	med.	large
inf-th	(34, 36)		
BMSC-BAYES	(11, 40), (15, 21)		
BMSC-RISS	(11, 40), (15, 20), (15, 21) (16, 20), (16, 21), (20, 21), (25, 41), (28, 32), (33, 34), (33, 35), (33, 36), (34, 35) (34, 36)	(30, 33)	(17, 19) (24, 25)
RUNS	(11, 40), (25, 41), (34, 36)		(23, 35) (22, 23)

(a) $M = \{m_0,\ m_1,\ m_2,\ m_3\}$

segments merged incorrectly		
merges involving		
small	med.	large
(34, 36)		
(11, 40), (15, 21)		
(11, 40), (15, 20), (15, 21) (16, 21), (20, 21), (25, 41), (28, 32), (33, 34), (33, 35), (33, 36), (34, 35), (34, 36)	(30, 33)	(17, 19), (24, 25)
(11, 40), (25, 41), (34, 36)		

(b) $M = \{m_1\}$

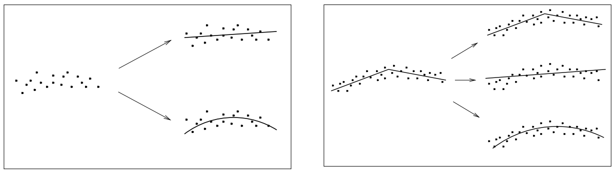
Table 7: Merging results for Image 2

criteria	19	24	37
AIC	14 x 14	22 x 22	14 x 14
CAIC	38 x 38	22 x 22	14 x 14
BIC	38 x 38	22 x 22	14 x 14
BAYES	38 x 38	22 x 22	14 x 14
RISS	70 x 70	26 x 26	22 x 22
BMSC-BAYES	89 x 89	38 x 38	22 x 22
BMSC-RISS	-	78 x 78	34 x 34
RUNS	10 x 10	14 x 14	10 x 10

Table 8: Model selection in small regions in segments 19, 24, and 37. Region size in pixels.

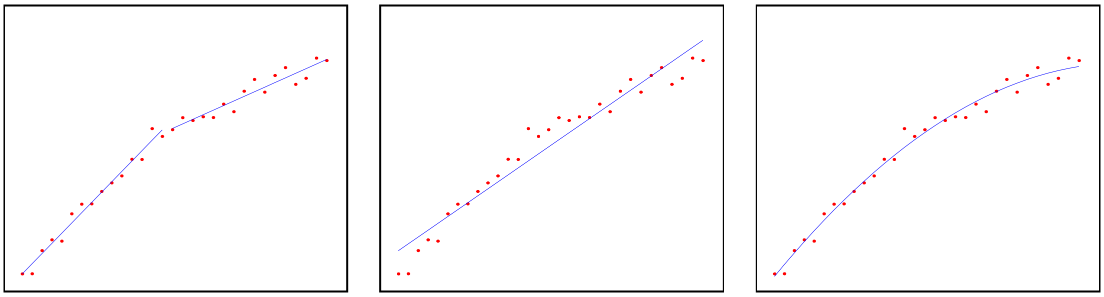
Performance	model selection	merging rules
Good	BAYES, BMSC-BAYES	BAYES, BIC, CAIC, BMSC-BAYES
Average	RUNS, CAIC, BIC	RUNS
Poor	RISS, BMSC-RISS, AIC	AIC, RISS, BMSC-RISS

Table 9: Overall performance for model selection and surface merging on Perceptron data.



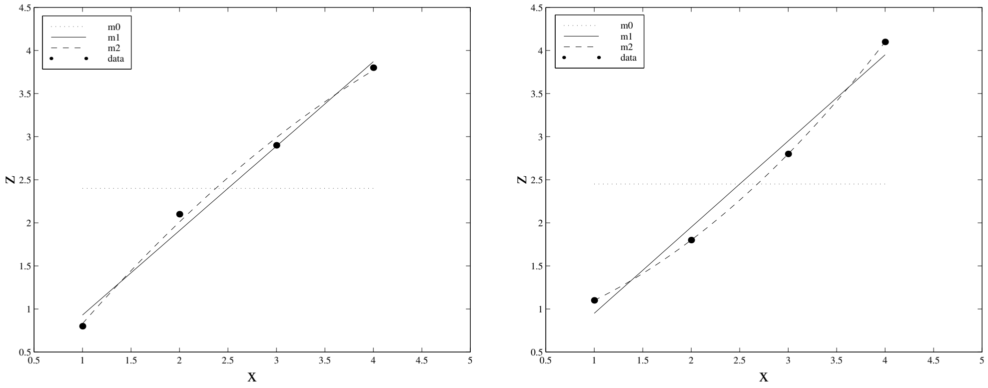
(a) (b)

Figure 1: (a) Model Selection: Determine the correct fitting function (model) to describe a data set; (b) Surface Merging: Given potentially oversegmented data, determine if the data should be represented by a single fit or by two different fits. The problem of model selection is implicit in merging.



(a) (b) (c)

Figure 2: Model selection and merging techniques can be used to determine the correct representation for the data even at small magnitude discontinuities.



(a) (b)

Figure 3: Shows two different samplings of the same true data points and fits corresponding to models m_0, m_1, and m_2. While fit accuracy remains almost the same for each model in the two samplings, the fit parameters change substantially for m_2 and remain stable for m_0 and m_1.

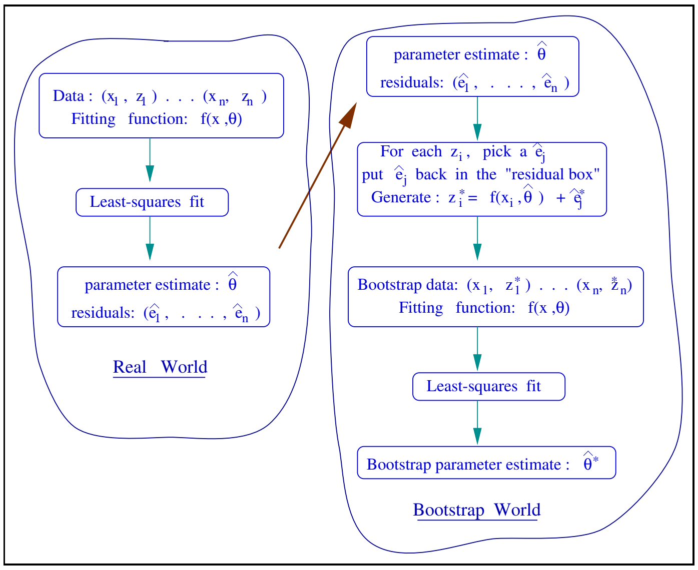
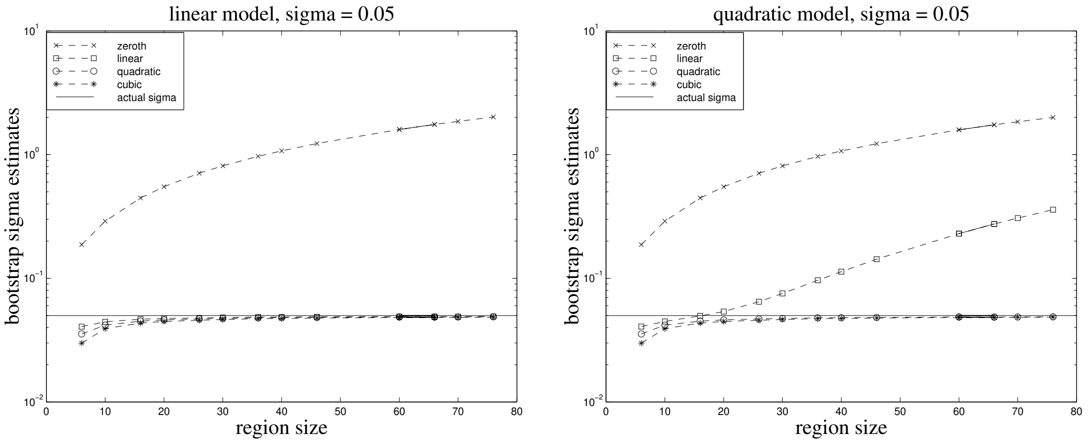
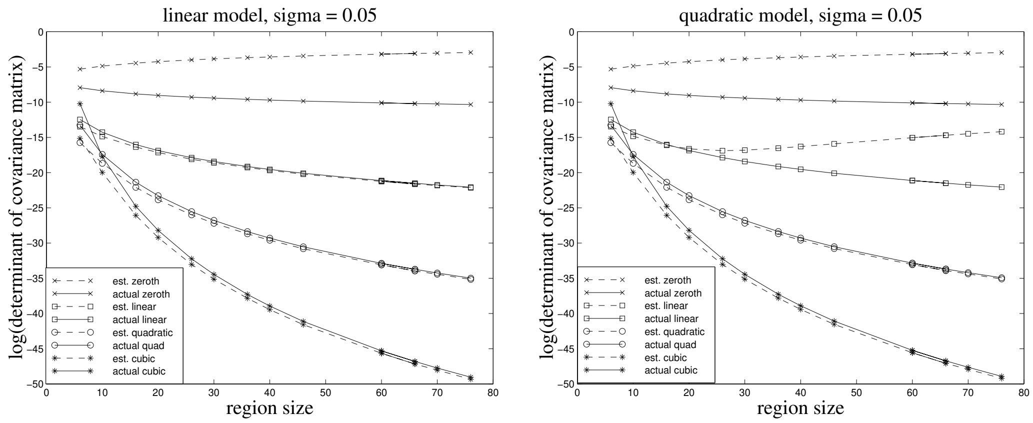


Figure 4: Schematic diagram of the bootstrap technique adapted from [19, chapter 8].



(a) (b)

Figure 5: Shows the bootstrap estimates of $\boldsymbol{\sigma}$ for data from linear and quadratic functions using different models. (a) shows the results for data generated from the linear model, and (b) shows the results for data generated from the quadratic model.



(a) (b)

Figure 6: Compares $\log|\boldsymbol{V}^*(\boldsymbol{\theta}_m)|$ with the expected value of $\log|\boldsymbol{V}(\boldsymbol{\theta}_m)|$ using different models at different region sizes.

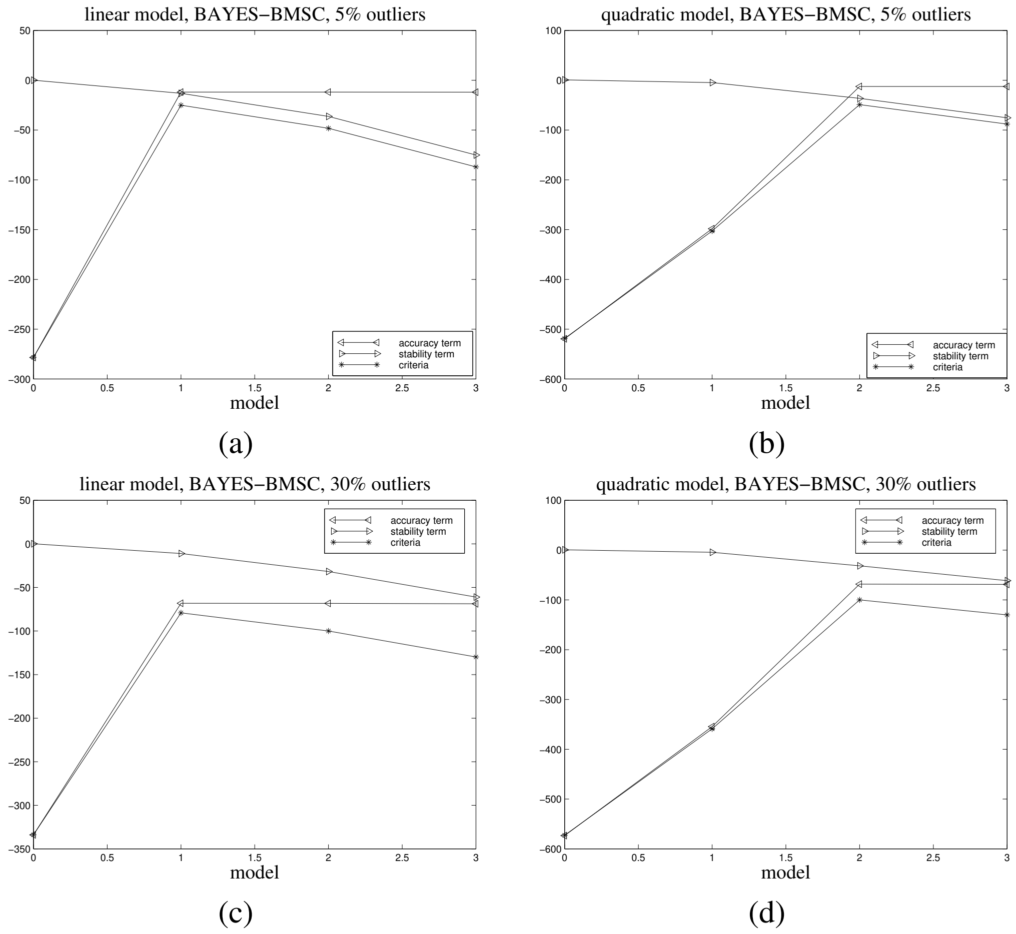


Figure 7: Shows the performance of BMSC-BAYES with different fractions of outliers in the data. The results are averaged over 50 simulations.

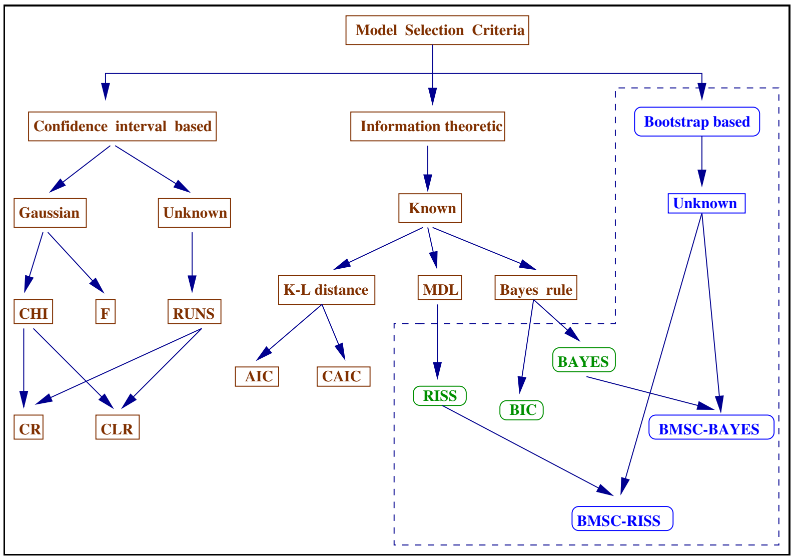


Figure 8: Classification of various model selection criteria presented in this paper. The criteria in the dashed box are either borrowed from the statistics literature or have been newly introduced in this paper. New merging rules on all of them are formulated in this paper.

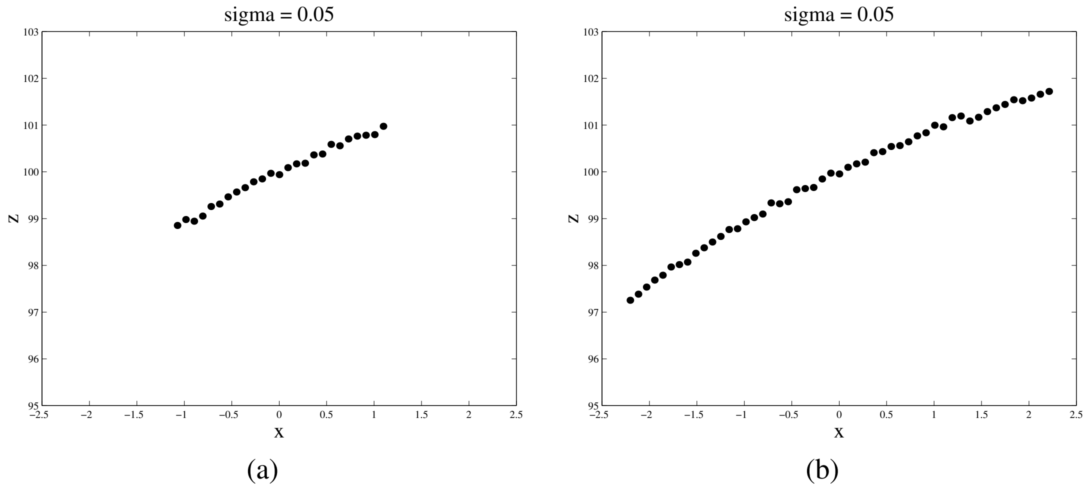
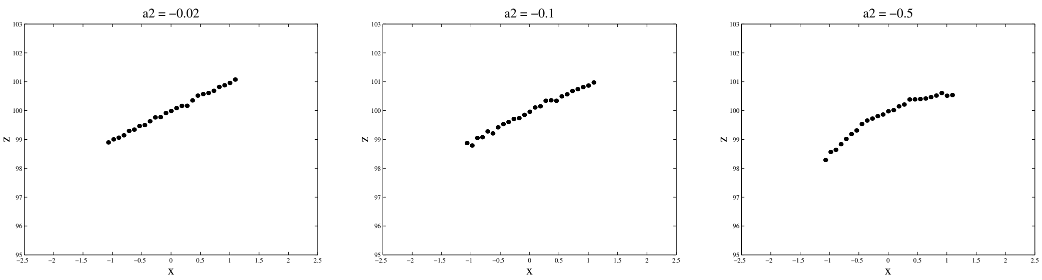
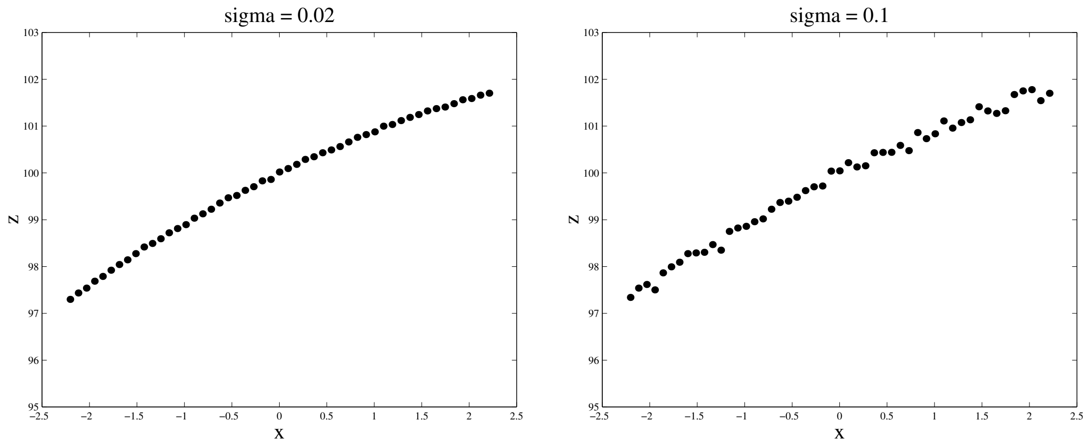


Figure 9: Sample data from a quadratic model at region sizes (a) 25 pixels and (b) 50 pixels at $\sigma = 0.05$.



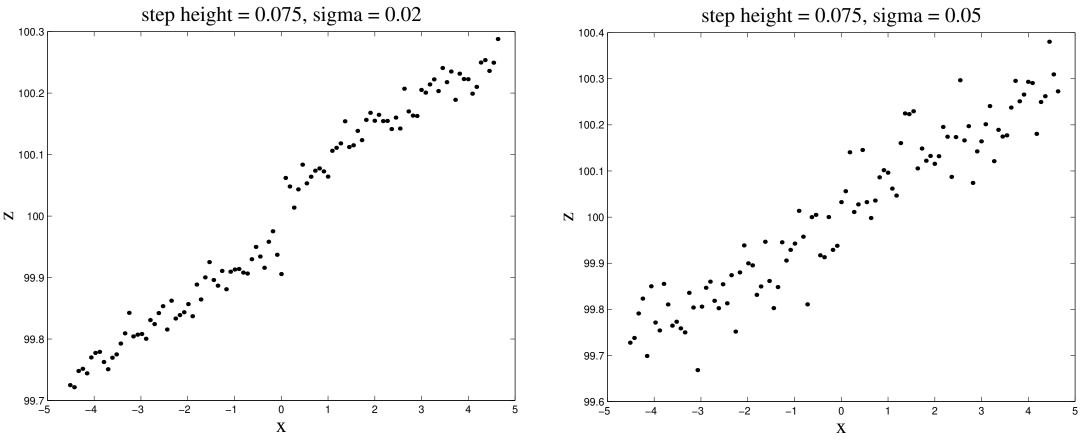
(a) (b) (c)

Figure 10: Shows sample data from a quadratic model with $\boldsymbol{a_0}$=100, $\boldsymbol{a_1}$=1, and (a) $\boldsymbol{a_2}$= -0.02, (b) $\boldsymbol{a_2}$= -0.1, and (c) $\boldsymbol{a_2}$=-0.5. Observe how the data in (a) and (b) appear to be from a linear model. All data points contain Gaussian errors with $\boldsymbol{\sigma = 0.05}$ cm.



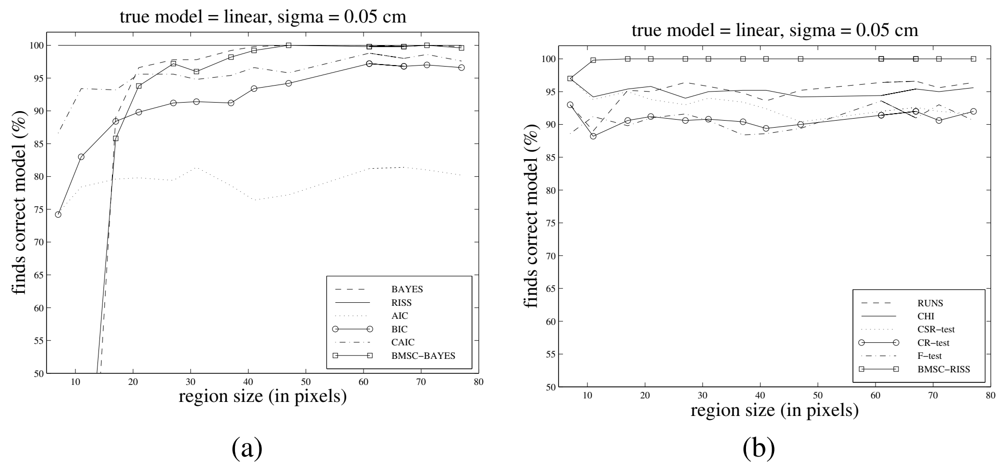
(a) (b)

Figure 11: Sample data from a quadratic model at $\boldsymbol{\sigma = 0.02}$ and $\boldsymbol{\sigma = 0.05}$.



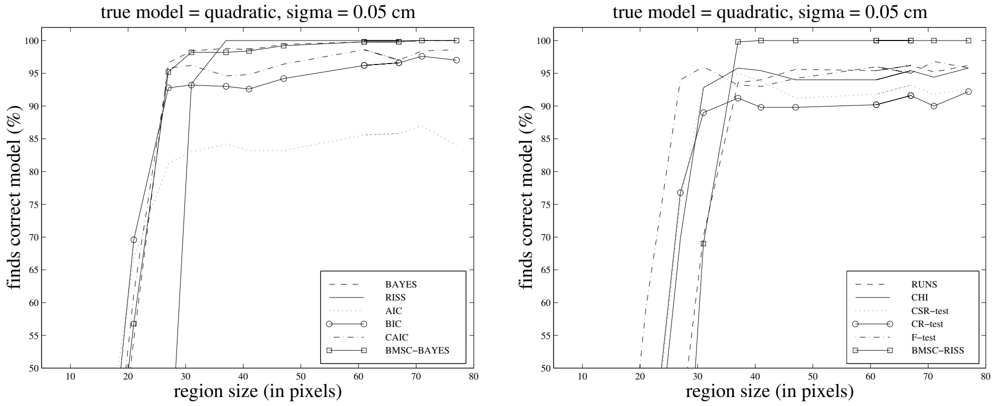
(a) (b)

Figure 12: Sample data from a step discontinuity at $x = 0$ with (a) $\sigma = 0.02$ and (b) $\sigma = 0.05$. The region size corresponding to each surface is 50 pixels. It is difficult to determine the discontinuity at $\sigma = 0.05$.



(a) (b)

Figure 13: Performance with increasing region size for data from linear model at $\sigma = 0.05$ cm. Data points are generated using Gaussian noise and the region size is increased symmetrically around the origin from 7 to 77 pixels.



(a) (b)

Figure 14: Model selection with changing region size for quadratic model at $\boldsymbol{\sigma = 0.05}$ cm.

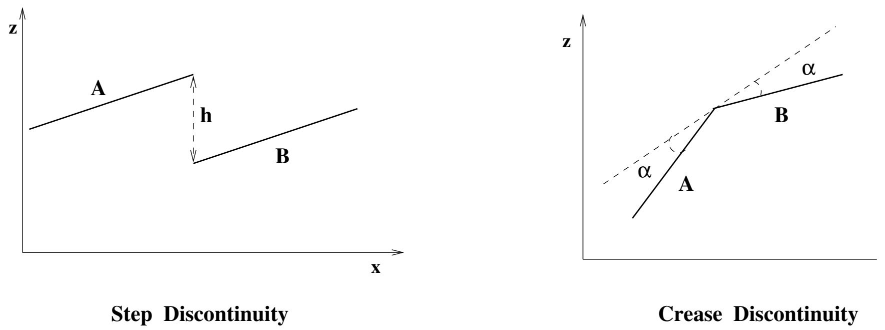
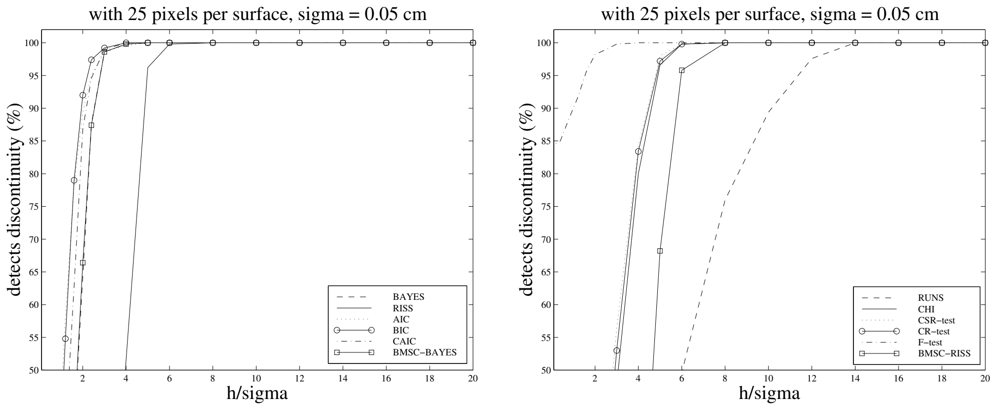
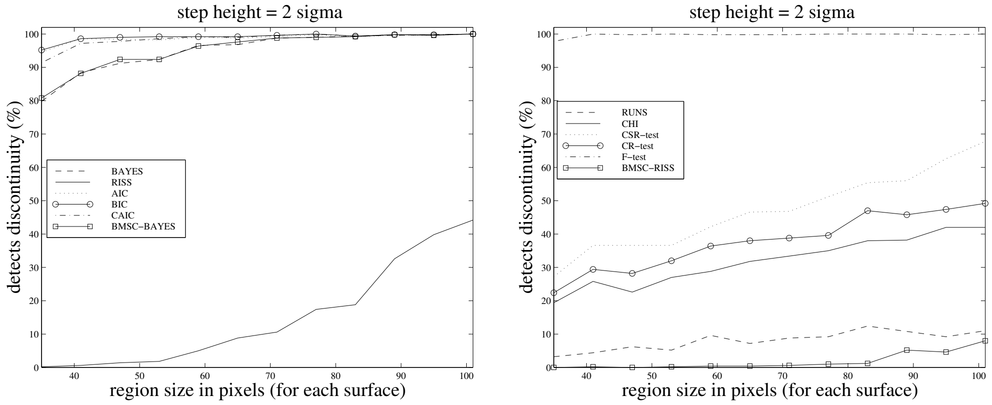


Figure 15: Shows step and crease discontinuity parameters.



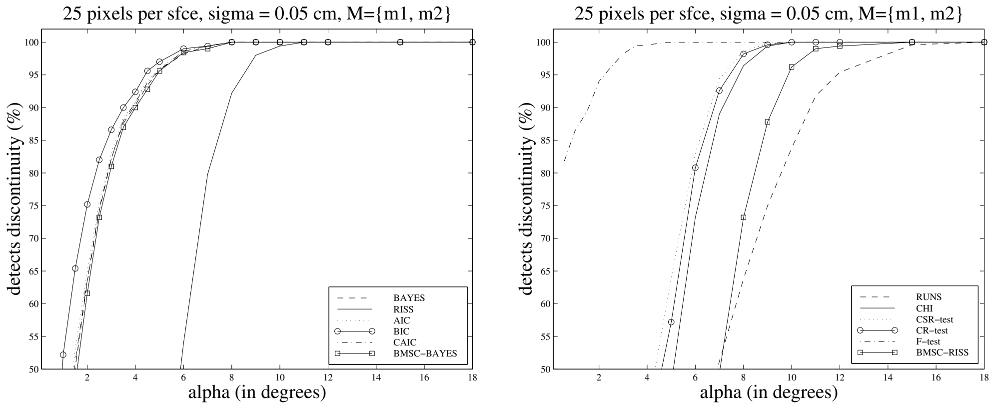
(a) (b)

Figure 16: Performance of merging rules at step discontinuities.



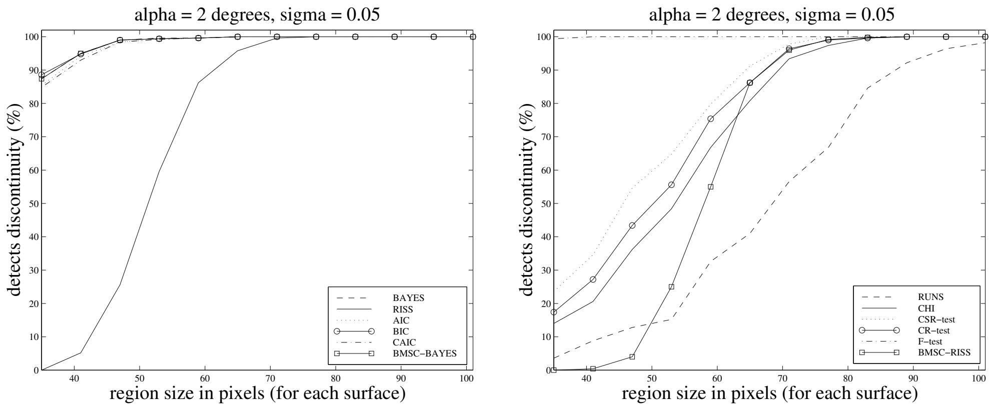
(a) (b)

Figure 17: Performance of merging rules at step discontinuities with increasing region sizes at $h = 2\sigma$.



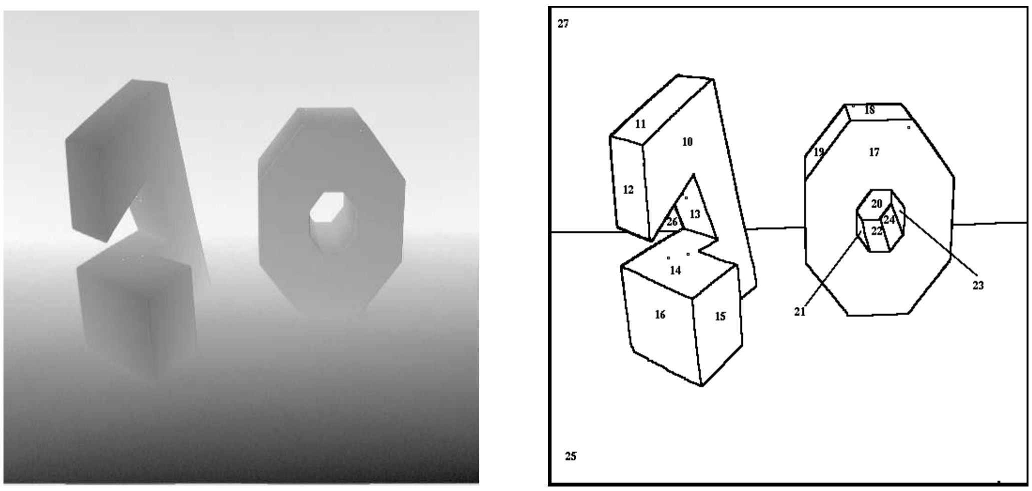
(a) (b)

Figure 18: Performance of merging rules at crease discontinuities with changing $\boldsymbol{\alpha}$.



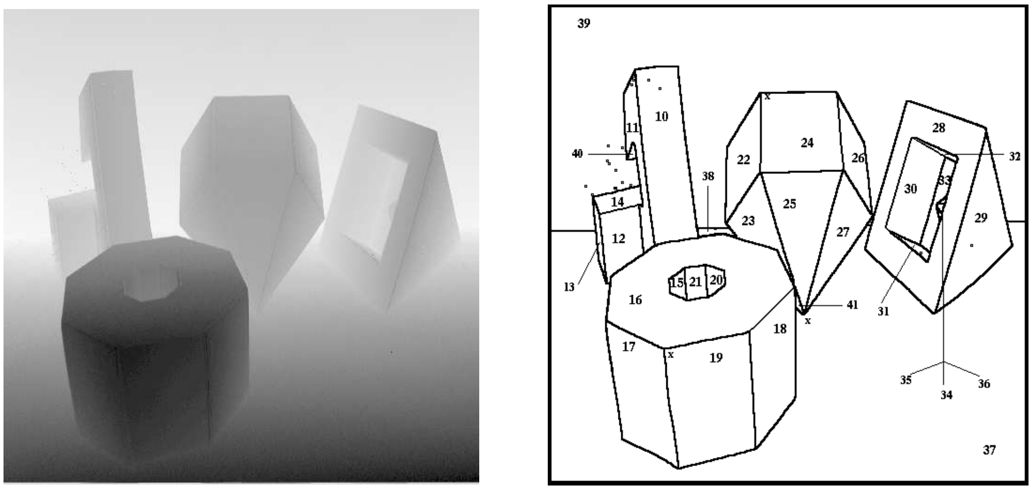
(a) (b)

Figure 19: Performance of merging rules at crease discontinuities with increasing region size at $\alpha = 2$ degrees.



(a) (b)

Figure 20: Model selection results for Image 1



(a) (b)

Figure 21: Model selection results for Image 2

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium of Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[2] F. Arman and J. K. Aggarwal. Model-based object recognition in dense-range images - a review. *ACM Computing Surveys*, 25(1):5–43, March 1993.

[3] R. H. Bartels and J. J. Jezioranski. Least-squares fitting using orthogonal multinomials. *ACM Transactions on Mathematical Software*, 11(3):201–217, Sept. 1985.

[4] P. J. Besl. *Surfaces in Range Image Understanding*. Springer-Verlag, 1988.

[5] P. J. Besl, J. B. Birch, and L. T. Watson. Robust window operators. In *ICCV*, pages 591–600, 1988.

[6] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE PAMI*, 10:167–192, 1988.

[7] R. M. Bolle and D. B. Cooper. Bayesian recognition of local 3-D shape by approximating image intensity functions with quadric polynomials. *IEEE PAMI*, 6(4):418–429, 1984.

[8] R. C. Bolles and M. A. Fischler. A RANSAC-based approach to model fitting and its applications to finding cylinders in range data. In *IJCAI*, pages 637–643, 1981.

[9] K. L. Boyer, M. J. Mirza, and G. Ganguly. The robust sequential estimator: A general approach and its application to surface organization in range data. *IEEE PAMI*, 16(10):987–1001, October 1994.

[10] H. Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

[11] K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*. John Wiley and Sons, Inc., 1960.

[12] K. Bubna. *Model Selection, Merging, and Splitting Techniques for Surface Reconstruction from Range Data*. PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, August 1998.

[13] K. Bubna and C. V. Stewart. Model selection and surface merging in reconstruction algorithms. In *ICCV*, pages 895–902, 1998.

[14] J. Cabrera and P. Meer. Unbiased estimation of ellipses by bootstrapping. *IEEE PAMI*, 18(7):752–756, 1996.

[15] F. S. Cohen and R. D. Rimey. A maximum likelihood approach to segmenting range data. In *IEEE Conference on Robotics and Automation*, pages 1696–1701, 1988.

[16] B. Curless and M. Levoy. Better optical triangulation through spacetime analysis. In *ICCV*, pages 987–993, Boston, MA, 1995.

[17] T. Darrell and A. Pentland. Cooperative robust estimation using layers of support. *IEEE PAMI*, 17(5):474–487, 1995.

[18] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley Publications, 1973.

[19] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[20] B. Finkenstadt, Q. Yao, and H. Tong. A conditional density approach to the order determination of time series. Technical Report UKC/IMS/96/17, Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent, UK, 1996.

[21] A. W. Fitzgibbon and R. B. Fisher. Lack-of-fit detection using the run-distribution test. In *European Conference on Computer Vision*, pages 173–178, Stockholm, 1994.

[22] F. Gustafsson and H. Hjalmarsson. Twenty-one ML estimators for model selection. *Automatica*, 31:1377–1392, October 1995.

[23] R. Hoffman and A. Jain. Segmentation and classification of range images. *IEEE PAMI*, 9:608–620, 1987.

[24] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.*, A6:813–827, 1977.

[25] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE PAMI*, 18:673–689, July 1996.

[26] E. T. Jaynes. *Probability Theory - the Logic of Science.* Physics, Washington University, St. Louis, MO 63130, USA, http://omega.albany.edu:8008/JaynesBook.html, 1994.

[27] S. M. LaValle and S. A. Hutchinson. A Bayesian segmentation methodology for parametric image models. *IEEE PAMI*, 17(2):211–217, Feb 1995.

[28] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. *IJCV*, 3:73–102, 1989.

[29] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *IJCV*, 14:253–277, 1995.

[30] M. Li. Minimum description length based 2D shape description. In *ICCV*, pages 512–517, 1993.

[31] J. A. F. Machado. Robust model selection and M-estimation. *Econometric Theory*, 9:478–493, 1993.

[32] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[33] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: A review. *IJCV*, 6:59–70, 1991.

[34] J. V. Miller and C. V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *CVPR*, pages 300–306, 1996.

[35] G. Qian and H. R. Kunsch. On model selection in robust linear regression. Technical Report 80, Seminar Fur Statistik, Eidgenossische Technische Hochschule (ETH), Zurich, Switzerland, Nov 1996.

[36] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:468–471, 1978.

[37] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[38] E. Ronchetti. Robust model selection in regression. *Statistical Probability Letters*, 3:21–23, 1985.

[39] B. Sabata, F. Arman, and J. K. Aggarwal. Segmentation of 3D range images using pyramidal data structures. *CVGIP:IU*, 57:373–387, 1993.

[40] K. Sato and S. Inokuchi. Range-imaging system utilizing nematic liquid crystal mask. In *ICCV*, pages 657–661, 1987.

[41] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[42] C. V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE PAMI*, 17(10):925–938, Oct. 1995.

[43] C. V. Stewart, K. Bubna, and A. Perera. Estimating model parameters and boundaries by minimizing a joint, robust objective function. In *CVPR*, 1999.

[44] C. V. Stewart, R. Y. Flatland, and K. Bubna. Geometric constraints and stereo disparity computation. *IJCV*, 20(3):143–168, 1996.

[45] J. Subrahmonia, D. B. Cooper, and D. Keren. Practical reliable Bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants. *IEEE PAMI*, 18(5):505–519, May 1996.

[46] F. S. Swed and C. Eisenhart. Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, 14:66–87, 1943.

[47] G. Taubin. Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range segmentation. *IEEE PAMI*, 13(11):1115–1138, 1991.

[48] R. Taylor, M. Savini, and A. Reeves. Fast Segmentation of Range Imagery Into Planar Regions. *CVGIP*, 45:42–60, 1989.

[49] P. H. S. Torr. An assesment of information criteria for motion model selection. In *CVPR*, pages 47–53, 1997.

[50] P. H. S. Torr, A. Fitzgibbon, and A. Zisserman. Maintaining multiple motion model hypotheses through many views to recover matching structure. In *ICCV*, pages 485–491, 1998.

[51] S. Weisberg. *Applied Linear Regression*. John Wiley and Sons, 1985.

[52] P. Whaite and F. P. Ferrie. Active exploration: knowing when we're wrong. In *ICCV*, pages 41–48, 1993.

[53] S. C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE PAMI*, 18(9):884–900, 1996.