Generalizing the Prediction Sum of Squares Statistic and Formula, Application to Linear Fractional Image Warp and Surface Fitting

Adrien Bartoli

Clermont Université, France

Adrien.Bartoli@gmail.com

September 8, 2016

Accepted in the *International Journal of Computer Vision*

Abstract

The Prediction Sum of Squares statistic uses the principle of Leave-One-Out Cross-Validation in Linear Least Squares regression. It is computationally attractive, as it can be computed non-iteratively. However, it has limitations: it does not handle coupled measurements, which should be held out simultaneously, and is specific to the principle of Leave-One-Out, which is known to overfit when used for selecting a model's complexity. We propose Multiple-Exclusion PRESS (MEXPRESS), which generalizes PRESS to coupled measurements and other types of Cross-Validation, while retaining computational efficiency with the non-iterative MEXPRESS formula. Using MEXPRESS, various strategies to resolve overfitting can be efficiently implemented. The core principle is to exclude training data too 'close' or too 'similar' to the validation data. We show that this allows one to select the number of control points automatically in three cases: *(i)* the estimation of linear fractional warps for dense image registration from point correspondences, *(ii)* surface reconstruction from a dense depth-map obtained by a depth sensor and *(iii)* surface reconstruction from a sparse point cloud obtained by Shape-from-Template.

Contents

1 Introduction

Cross-Validation (CV) is a useful principle to compare models and estimate hyper-parameters in regression. It has been used in many different types of problems in statistical learning (Bishop, 1995; Hastie et al., 2001; Schölkopf and Smola, 2001) and computer vision (Forsyth and Ponce, 2003), including problems in super-resolution (Nguyen et al., 2001), visual recognition (Jurie and Triggs, 2005), quantification of the perceptual image quality (Tang et al., 2011) and tracking (Sevilla-Lara and Learned-Miller, 2012). In Linear Least Squares (LLS) regression, the concept of CV is more specifically known as Prediction Sum of Squares (PRESS) (Allen, 1971). PRESS assumes that the fitting residuals are normal and IID, and can thus only be computed for noisy datasets which do not contain blunders. We bring a set of new results on the computation of PRESS. We apply these results in correspondence-based dense image registration, substantially extending our previous work (Bartoli, 2008, 2009) on linear warps to the class of linear fractional warps modeling perspective projection. We also apply these results in surface fitting with dense and sparse data, obtained from two different sources, namely a depth sensor and Shape-from-Template (SfT).

CV produces a score quantifying the predictive ability or *predictivity* of a model with respect to a *dataset*, independently of the actual model parameters. The basic principle is to split the data in two subsets: the *training set* and the *validation set*. The model is fitted to the training set and its predictivity is then measured on the validation set. Most CV scores average predictivity over multiple training and validation splits of the data. One of the most popular types of CV is Leave-One-Out CV (LOOCV), which uses each of the n data in turn as validation sets and the remaining $n-1$ data as training sets. Computing the LOOCV score may be expensive as directly applying the basic principle requires one to fit the model n times. Fortunately, this is not the case in LLS regression as the LOOCV score corresponds to the PRESS statistic which may be computed by the non-iterative PRESS formula (Yan and Su, 2009). Non-iterative means that the model does not have to be fitted repeatedly but only once, to the complete dataset. The PRESS formula is thus extremely interesting from a computational stand-point. It has however two main limitations. First, it is specific to LOOCV. Second, it makes the one-datum-one-measurement hypothesis. In the dataset each datum corresponds to a physical entity, such as a point correspondence in image registration. A measurement however is represented by a single 'equation' or 'constraint'. The one-datum-one-measurement hypothesis holds if the dataset is in one-to-one correspondence with the *measurement set*. In other words, it holds if a datum provides only one measurement. However, a datum typically provides several measurements, one for each of the g problem's dimensions, and we therefore have $m = gn$ measurements organized in groups of g *coupled measurements*. In 2D image registration, a point correspondence provides two measurements, and for n point correspondences we thus have $m = 2n$ measurements (Hartley and Zisserman, 2003, Chapter

4). The one-datum-one-measurement hypothesis typically holds in one-dimensional problems such as curve fitting but may break in higher-dimensional problems such as correspondence-based image registration and 3D transformation fitting. In the PRESS statistic the measurements are held out individually, whereas they should be held out in groups of size g corresponding to each datum to compute the LOOCV score. In some special cases discussed in §2.2 however, the PRESS statistic may still give the LOOCV score in higher dimensions.

We propose the Multiple-Exclusion PRESS (MEXPRESS) formula as a tool to generalize the PRESS formula. Our goal is to handle models for which the one-datum-one-measurement hypothesis does not hold and to compute PRESS statistics for other types of CV scores than LOOCV, all of them non-iteratively. The MEXPRESS formula computes a model's predictivity measured on any combination of training and validation sets obtained by splitting the data, non-iteratively. In other words, the MEXPRESS formula does not require one to fit the model specifically to the training set at hand. The MEXPRESS formula computes the prediction residuals for any subset $\mathcal{K} \subset [1, m]$ of measurements given the residuals obtained by fitting the model to all measurements, similarly to the PRESS formula, and is specific to LLS regression. The basic use of the MEXPRESS formula is to compute the Leave-One-Out PRESS statistic non-iteratively when the one-datum-one-measurement assumption does not hold. The MEXPRESS formula can also be used to compute other forms of PRESS statistics non-iteratively, such as the k-Fold, Leave-p-Out and Random-Sampling PRESS statistics, whether the one-datum-one-measurement hypothesis holds or not. LOOCV (and thus the PRESS statistic) is an almost unbiased estimator for the prediction error but may have high variance, while k-Fold CV has lower variance but may be significantly biased (Hastie et al., 2001). We propose a new type of CV score, the Local-Exclusion CV score, which mitigates both effects, and instantiate it as the Local-Exclusion PRESS statistic for LLS problems. Local-Exclusion CV resembles LOOCV as it averages the model's predictivity over n singleton validation sets formed by each of the n data. However, it differs in that the model is not trained on the remaining $n-1$ data: the training sets contain at most $n-1$ data but may be smaller, as they exclude the data too close or too similar to the validation datum. The notions of closeness and similarity have to be interpreted in a problem specific way. Using the MEXPRESS formula, the Local-Exclusion PRESS statistic can be computed non-iteratively and does not require the one-datum-one-measurement hypothesis to hold. We use the MEXPRESS formula to explicitly give three PRESS formulas: the Coupled-Measurements, the k-Fold and the Local-Exclusion PRESS formulas, which compute the corresponding PRESS statistics.

The proposed PRESS formulas can be used in correspondence-based dense image registration. The task is to estimate a continuous smooth function mapping points from a source to a target image from n point correspondences. The function to be estimated is called image warp or simply *warp* and is often represented by a parametric model, typically a Thin-Plate Spline (TPS) (Bookstein, 1989) or a Bicubic B-Spline (BBS) (Rueckert et al., 1999), which lie in the wider class of linear warps, and may be estimated by LLS regression. The warp's optimal number of control points may be selected automatically by LOOCV (Bartoli, 2008, 2009). We are interested in the computation of the class of *linear fractional warps* such as the Deformable-Perspective warp constructed from the TPS (Bartoli et al., 2010) and the NURBS warp constructed from the BBS (Brunet et al., 2009). Linear fractional warps form a particularly important class as the fractional part models perspective projection. The basic fitting principle is to transfer each of the n source image points to the target image by applying the unknown warp and minimize their discrepancy to the corresponding target image points. Though the linear fractional warps are nonlinear, they may be estimated by LLS regression using the algebraic distance (Hartley and Zisserman, 2003, Chapter 4). Each point correspondence yields $g = 2$ measurements. We show how the MEXPRESS formula allows one to compute the LOO, k-Fold and Local-Exclusion CV scores non-iteratively for a linear fractional warp. We then use these scores to select the warp's number of control points automatically. The proposed PRESS formulas can also be used in surface fitting. The task is to estimate a continuous smooth function embedding a subset of the real plane to a surface in space. We considered two types of data from which this function is to be estimated. The first type is dense depth-maps, obtained from a depth sensor. The second type is sparse point clouds, obtained by SfT. In both cases, the function is represented by a parametric linear model, and has control points placed automatically by a data-adaptive strategy inspired by similar strategies from curve-fitting (Dierckx, 1981, 1993). We show how the MEXPRESS formula allows one to compute the LOO, k-Fold and Local-Exclusion CV scores non-iteratively for the surface functions for both types of data, namely the depth and the embedding functions. We then use these scores to select the function's number of control points automatically. As expected, the higher the number of control points, the lower the fitting residual, which therefore cannot be used to select the number of control points. Using a PRESS statistic to select the number of control points is also motivated by the fact that one cannot easily compute a measure of statistical significance required to use, for instance, the fitting residual as selection criterion. The reason is twofold and grounded in the conditions holding for the vast majority of practical cases where automatically selecting the number of control points will be required. First, the level of noise and modeling error are unknown. Second, the amount of data is limited, making the computation of spread statistics such as the standard deviation unreliable. For these reasons, one preferably requires criteria whose graph gives the number of control points to select as one of its minima. Our experiments suggest that the new PRESS statistic using an exclusion radius in its internal training stage is, on the one hand, substantially less sensitive to overfitting than the existing criteria and, on the other hand, does not underfit either.

Notation and terminology. We use regular fonts in italics for scalars such as $E \in \mathbb{R}$ and in bold for vectors such as $\mathbf{x} \in \mathbb{R}^p$, type writer fonts for matrices such as $\mathtt{A} \in \mathbb{R}^{m \times p}$ and calligraphic fonts for sets such as $\mathcal{K} \subset \mathbb{N}$. The size of a set is written as $|\mathcal{K}|$. Matrix tranpose, inverse and pseudo-inverse are respectively written as in $\mathtt{A}^\top$, $\mathtt{A}^{-1}$ and $\mathtt{A}^\dagger \overset{\text{def}}{=} \left(\mathtt{A}^\top \mathtt{A}\right)^{-1} \mathtt{A}^\top$. The identity matrix is written as $\mathtt{I}$. We use the stack operator $\text{stk}(\mathbf{x}, \mathbf{y}, \dots) \overset{\text{def}}{=} [\mathbf{x}^\top \; \mathbf{y}^\top \; \cdots]^\top$. The operators to select the elements respectively indicated and not indicated in $\mathcal{K}$ in a vector $\mathbf{x}$ or the rows in a matrix $\mathtt{A}$ are $\mathbf{x}_{\mathcal{K}}$ and $\mathbf{x}_{-\mathcal{K}}$. We use the standard notation $\mathbf{x}_{(\mathcal{K})}$ to mean 'done without' the elements indicated in $\mathcal{K}$. If $\mathcal{K} = \{j\}$ contains just one element we may simply write $\mathbf{x}_j$, $\mathbf{x}_{-j}$ and $\mathbf{x}_{(j)}$. We define $\mathtt{A}_{\mathcal{K},\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{K}|}$ as the submatrix formed with the $|\mathcal{K}|$ rows and columns of $\mathtt{A}$ with index in $\mathcal{K}$. We define $\mathtt{C}_{f,r} \in \mathbb{R}^{g \times gr}$ as the block-wise row matrix containing $r-1$ blocks $\mathtt{0} \in \mathbb{R}^{g \times g}$ and one block $\mathtt{I} \in \mathbb{R}^{g \times g}$ as its fth block. The LLS problem is formulated from n data. We assume for notation simplicity that each datum gives g measurements so the total number of measurements is $m = gn$, but our results also hold when the data do not all give the same number of measurements. We define $\mathtt{1}$ as the 'all-one' matrix. We use $[\mathbf{u}]_\times \mathbf{v} \overset{\text{def}}{=} \mathbf{u} \times \mathbf{v}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, where $[\mathbf{u}]_\times \in \mathbb{R}^{3 \times 3}$ is the skew-symmetric cross-product matrix. We use $\|\mathbf{x}\|_2$ and $\|\mathtt{A}\|_{\mathcal{F}}$ for the vector two-norm and the matrix Frobenius norm, respectively.

2 Background

2.1 Linear Least Squares Regression

The model's parameters are held by the parameter vector $\mathbf{x} \in \mathbb{R}^p$. Each of the m measurements is a pair formed by a vector $\mathbf{a}_j \in \mathbb{R}^p$ called regressor vector and a scalar $b_j \in \mathbb{R}$ called response, with $j = 1, \dots, m$. The regressor vectors form the rows of the matrix $\mathtt{A} \overset{\text{def}}{=} [\mathbf{a}_1 \; \cdots \; \mathbf{a}_m]^\top \in \mathbb{R}^{m \times p}$ called the design matrix and the responses form the elements of the vector $\mathbf{b} \in \mathbb{R}^m$ called the response vector. The model with parameters $\mathbf{x}$ fits the response vector as $\mathtt{A}\mathbf{x}$ and its least squares estimate is given by $\bar{\mathbf{x}} \overset{\text{def}}{=} \mathtt{A}^\dagger \mathbf{b}$. The least squares estimate minimizes the Root Mean Square (fitting) Residual (RMSR) E defined as:

$$E^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{m} e_j^2 = \frac{1}{n} \sum_{j=1}^{m} \left(b_j - \mathbf{a}_j^\top \bar{\mathbf{x}}\right)^2 = \frac{1}{n} \|\mathbf{b} - \mathtt{A}\bar{\mathbf{x}}\|_2^2 = \frac{1}{n} \left\|\left(\mathtt{I} - \hat{\mathtt{A}}\right) \mathbf{b}\right\|_2^2, \tag{1}$$

where $\hat{\mathtt{A}} \overset{\text{def}}{=} \mathtt{A}\mathtt{A}^\dagger \in \mathbb{R}^{m \times m}$ is the hat matrix. The average is taken for the number n of data rather than for the number m of measurements. LLS regression naturally handles coupled measurements. This is because the RMSR can be rescaled without changing the estimate.

2.2 The Prediction Sum of Squares Statistic

The PRESS statistic follows the LOOCV principle and is thus an exhaustive statistic. It has been formulated under the one-datum-one-measurement hypothesis (Allen, 1971; Yan and Su, 2009) and therefore does not handle coupled measurements. Each response b_j is predicted as $\mathbf{a}_j^\top \bar{\mathbf{x}}_{(j)}$, where $\bar{\mathbf{x}}_{(j)} \stackrel{\text{def}}{=} (\mathtt{A}_{-j})^\dagger \mathbf{b}_{-j}$ are the parameters obtained by fitting the model without the jth measurement. The PRESS statistic is then given by:

$$P^2 \ \stackrel{\text{def}}{=}\ \frac{1}{m}\sum_{j=1}^{m} e_{(j)}^2 \ = \ \frac{1}{m}\sum_{j=1}^{m}\left(b_j - \mathbf{a}_j^\top \bar{\mathbf{x}}_{(j)}\right)^2. \tag{2}$$

The PRESS formula (Yan and Su, 2009) allows one to compute the PRESS statistic non-iteratively as:

$$P^2 \ = \ \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\left(1-\hat{a}_{j,j}\right)^2} e_j^2 \ = \ \frac{1}{m}\left\|\operatorname{diag}\left(\frac{1}{1-\hat{a}_{1,1}},\cdots,\frac{1}{1-\hat{a}_{m,m}}\right)\left(\mathtt{I}-\hat{\mathtt{A}}\right)\mathbf{b}\right\|_2^2. \tag{3}$$

The average is taken for the number m of measurements. The PRESS formula (3) was derived under the one-datum-one-measurement hypothesis, which implies $g = 1$ and thus $m = n$. It does not hold in the presence of coupled measurements caused by $g \neq 1$. This is because the measurements are being held out individually to form the PRESS statistic (2), whereas they should be held out in groups of size g corresponding to each datum. Therefore, the PRESS statistic (2) almost always underestimates the ‘true’ PRESS statistic for $g \neq 1$. This case occurs very commonly as it corresponds to models with multiple dimensions. We extended the PRESS formula to cope with $g \neq 1$ in the special case where the model's multiple dimensions share their regressor vectors (Bartoli, 2009). In other words, the model's parameters must be organized in a matrix $\mathtt{X} \in \mathbb{R}^{\frac{p}{g}\times g}$ rather than in a vector $\mathbf{x} \in \mathbb{R}^p$. Each column of matrix $\mathtt{X}$ gives the model's parameter for one dimension. The design matrix $\mathtt{A} \stackrel{\text{def}}{=} [\mathbf{a}_1 \cdots \mathbf{a}_n]^\top \in \mathbb{R}^{n\times\frac{p}{g}}$ holds only one copy of the regressor vector per datum. Matrix $\mathtt{B} \stackrel{\text{def}}{=} [\mathbf{b}_1 \cdots \mathbf{b}_n]^\top \in \mathbb{R}^{n\times g}$ contains the responses with each of its columns corresponding to one of the problem's dimensions. The RMSR is given by $E^2 = \frac{1}{n}\|\mathtt{A}\bar{\mathtt{X}} - \mathtt{B}\|_{\mathcal{F}}^2$ and the least squares estimate by $\bar{\mathtt{X}} = \mathtt{A}^\dagger\mathtt{B}$. Defining the per-datum fitting residual as $\mathbf{e}_i^\top \stackrel{\text{def}}{=} \mathbf{b}_i^\top - \mathbf{a}_i^\top\bar{\mathtt{X}}$ the PRESS statistic is given by:

$$P^2 \ = \ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left(1-\hat{a}_{i,i}\right)^2}\|\mathbf{e}_i\|_2^2 \ = \ \frac{1}{n}\left\|\operatorname{diag}\left(\frac{1}{1-\hat{a}_{1,1}},\cdots,\frac{1}{1-\hat{a}_{n,n}}\right)\left(\mathtt{I}-\hat{\mathtt{A}}\right)\mathtt{B}\right\|_{\mathcal{F}}^2. \tag{4}$$

There has been no PRESS formula proposed to cope with the general setup of coupled measurements when $g \neq 1$, and to other types of CV than LOOCV such as k-Fold CV. We propose MEXPRESS and the MEXPRESS formula which allow us to solve these computations efficiently. MEXPRESS has the results we obtained in (Bartoli, 2009) as special cases.

3 Generalizing the PRESS Statistic and Formula

Our first goal is to give general means to easily compute PRESS non-iteratively. We achieve this goal by providing the MEXPRESS formula. Our second goal is to extend the PRESS formula to compute the 'true' PRESS statistic for the general case of coupled measurements and coupled responses. Our third goal is to extend the PRESS statistic and formula to other types of CV such as k-Fold CV. In particular we want to be able to exclude data in the held out phase to reduce LOOCV's overfitting. Overfitting occurs when the selected model is excessively complex and leads to bad predictive performance. We achieve this goal with a novel PRESS statistic called Local-Exclusion PRESS.

3.1 The Basic Tool: MEXPRESS

The Multiple-Exclusion PRESS (MEXPRESS) formula computes the prediction residual $\mathbf{e}_{(\mathcal{K})}$ of multiple measurements with index set $\mathcal{K} \subset [1, m]$ by holding them out in model fitting:

$$\mathbf{e}_{(\mathcal{K})} \overset{\text{def}}{=} \mathbf{b}_{\mathcal{K}} - \mathtt{A}_{\mathcal{K}} \bar{\mathbf{x}}_{(\mathcal{K})}, \tag{5}$$

where $\bar{\mathbf{x}}_{(\mathcal{K})}$ is the model parameters fitted without using the measurements in $\mathcal{K}$. Proposition 1 gives the MEXPRESS formula. Its proof is provided in Appendix A. The MEXPRESS formula allows one to compute $\mathbf{e}_{(\mathcal{K})}$ independently of the model parameters $\bar{\mathbf{x}}_{(\mathcal{K})}$ and only as a function of the global model estimate $\bar{\mathbf{x}}$ and the hat matrix $\hat{\mathtt{A}}$.

Proposition 1. *The MEXPRESS formula establishes that:*

$$\mathbf{e}_{(\mathcal{K})} = \left(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\right)^{-1} \mathbf{e}_{\mathcal{K}}, \tag{6}$$

where $\mathtt{I} \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{K}|}$ *is an identity matrix,* $\hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{K}|}$ *is a submatrix formed with the* $|\mathcal{K}|$ *rows and columns of* $\hat{\mathtt{A}}$ *with index in* $\mathcal{K}$ *and* $\mathbf{e}_{\mathcal{K}} = \mathtt{A}_{\mathcal{K}} \mathtt{A}^{\dagger} \mathbf{b} \in \mathbb{R}^{|\mathcal{K}|}$ *are the global fit residuals.*

We observe that the MEXPRESS formula generalizes the equality $e_{(j)} = \frac{1}{1-\hat{a}_{j,j}} e_j$ which lies at the heart of the PRESS formula (3). The measurement index j is replaced by the measurement set $\mathcal{K}$, and scalar inverse by matrix inverse.

We define a partitionwise PRESS statistic as a PRESS statistic computed for measurements partitioned in t groups $\mathcal{K}_1, \ldots, \mathcal{K}_t$ to be held out jointly. Each group is a set which holds the indices of one or several measurements. The partition constraint means that the union of all groups covers all measurements, that a measurement is held by exactly one group, and that a group is non-empty. A partitionwise PRESS statistic is defined as:

$$P^2(\mathcal{K}_1,\dots,\mathcal{K}_t) \;=\; \frac{1}{n}\sum_{w=1}^{t}\left\|\mathbf{e}_{(\mathcal{K}_w)}\right\|_2^2 \;=\; \frac{1}{n}\sum_{w=1}^{t}\left\|\mathbf{b}_{\mathcal{K}_w}-\mathtt{A}_{\mathcal{K}_w}\bar{\mathbf{x}}_{(\mathcal{K}_w)}\right\|_2^2. \tag{7}$$

A non-iterative formula for a partitionwise PRESS statistic is obtained by applying proposition 1 to equation (7):

$$P^2(\mathcal{K}_1,\dots,\mathcal{K}_t) \;=\; \frac{1}{n}\sum_{w=1}^{t}\left\|\left(\mathtt{I}-\hat{\mathtt{A}}_{\mathcal{K}_w,\mathcal{K}_w}\right)^{-1}\mathbf{e}_{\mathcal{K}_w}\right\|_2^2 \tag{8}$$

$$\;=\; \frac{1}{n}\left\|\operatorname{diag}\left(\left(\mathtt{I}-\hat{\mathtt{A}}_{\mathcal{K}_1,\mathcal{K}_1}\right)^{-1},\cdots,\left(\mathtt{I}-\hat{\mathtt{A}}_{\mathcal{K}_t,\mathcal{K}_t}\right)^{-1}\right)\left(\mathtt{I}-\hat{\mathtt{A}}\right)\mathbf{b}\right\|_2^2. \tag{9}$$

The involved identity matrices are generally not of the same size, and the hat matrix is in general not block diagonal. However, only its diagonal blocks are used to weight the global fit residuals. An efficient implementation of equation (9) may thus first compute the hat matrix $\hat{\mathtt{A}}$ and then extract and inverse its diagonal blocks, which are typically of size much smaller than the total number of measurements.

3.2 Handling Coupled Measurements and Coupled Responses

An example of problem where each datum gives more than one measurement is when fitting an image warp to point correspondences, as each correspondence gives two coupled measurements, as studied in §4. The m measurements are partitioned into n groups $\mathcal{J}_1,\dots,\mathcal{J}_n$ of size g with $ng=m$, where each group corresponds to a datum. Assuming that the measurements are ordered per datum, we have:

$$\mathcal{J}_i \;\stackrel{\text{def}}{=}\; [g(i-1)+1, gi] \quad \text{for } i=1,\dots,n. \tag{10}$$

The RMSR is defined as an average over the number of data rather than the number of measurements:

$$E^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{b}_{\mathcal{J}_i}-\mathtt{A}_{\mathcal{J}_i}\bar{\mathbf{x}}\|_2^2 \;=\; \frac{1}{n}\|\mathbf{b}-\mathtt{A}\bar{\mathbf{x}}\|_2^2. \tag{11}$$

We define the PRESS statistic for this coupled measurements setup and derive its non-iterative formula from the partitionwise PRESS in equations (7,9) respectively, as $P^2(\mathcal{J}_1,\dots,\mathcal{J}_n)$. Its implementation has to compute the hat matrix $\hat{\mathtt{A}}$ and then inverse its diagonal blocks of size $g\times g$. The complexity of its computation depends on the group size g to compute the first factor, and the total number of measurements m to compute the second factor. In practice however, we have $g\ll m$, and the complexity is dominated by the computation of the second factor, which is the same as solving standard LLS regression. When fitting a warp or a global coordinate transform we typically have $g\in\{2,3\}$.

MEXPRESS naturally handles coupled responses. These correspond to measurements which share their regressor vectors but whose responses are different. Physically, they correspond to repeated observations with the same system state. These measurements must be held out simultaneously to compute PRESS. They can thus be grouped by gathering their indices in some set $\mathcal{K}$. The MEXPRESS formula is then very similar to the PRESS formula (9) but takes a special form as $\hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}} \propto \mathtt{1}$, the 'all-one' matrix.

3.3 Using k-Fold Rather than Leave-One-Out

The k-Fold PRESS formula implements the non-exhaustive k-Fold CV non-iteratively. Computing k-Fold PRESS using MEXPRESS is very similar to computing Coupled-Measurements PRESS. The main difference is that in k-Fold PRESS the m measurements are partitioned into fewer k groups $\mathcal{H}_1, \dots, \mathcal{H}_k$ of larger size s with typically $k \in \{5, 10\}$ and $s = \frac{m}{k}$ (though because the sizes of measurement groups must sum to m some groups will have size $\frac{m}{k}$ and some others will have size $\frac{m}{k} + 1$). The k-Fold PRESS statistic is thus given by the partitionwise PRESS (7) as $P^2(\mathcal{H}_1, \dots, \mathcal{H}_k)$. Its non-iterative formula is then directly given by equation (9). The resulting formula does not require the one-datum-one-measurement hypothesis to compute the k-Fold PRESS, as the leading factor $\frac{1}{n}$ suggests. If a coupling pattern exists between the measurements, they must simply be included in the same group to be held out jointly. If the one-datum-one-measurement hypothesis holds, then $g = 1$ and $n = m$, and the k-Fold PRESS formula simply considers no coupling between the measurements.

3.4 Handling Data Dependencies with Local Exclusion

We propose a novel exhaustive PRESS statistic that mitigates the overfitting effect of the PRESS statistic, and its non-iterative computation. The key idea is that some measurements may be highly correlated, typically when the measurement set is large. For instance when estimating an image warp, if the number of correspondences is large, then computing the PRESS statistic by holding out one correspondence at a time and using the PRESS formula (9) will presumably not be helpful in selecting the model's hyper-parameters, as many of the most complex models will fit the data equally well, and the selected model may overfit the data. In Local-Exclusion PRESS, we propose to hold out several measurements jointly, as we did to handle coupled measurements and in k-Fold PRESS, but with three fundamental differences, which are *(i)* to choose the groups of measurements to avoid overfitting based on a notion of closeness or similarity between the data driven by the problem at hand, *(ii)* to measure the prediction residual for only one datum at a time and *(iii)* to use overlapping measurement groups to ensure the statistic's exhaustivity. In the example of warp fitting this means holding out each correspondence at a time with its neighbors, fitting the warp, and computing the prediction residual only at the correspondence at hand (and not at the correspondence's neighbors). Local-Exclusion PRESS thus requires one to define a data distance function $\delta : \mathbb{N}^2 \to \mathbb{R}$ and the neighborhood size $\tau \in \mathbb{R}$. We define n possibly overlapping groups $\mathcal{K}_1, \dots, \mathcal{K}_n$ with:

$$\mathcal{K}_i \overset{\text{def}}{=} \{\mathcal{J}_{i'} \,|\, i' \in [1, n] \wedge \delta(i, i') \leq \tau\} \text{ for } i = 1, \dots, n, \tag{12}$$

where $\mathcal{J}_i$ indicates the set of measurements related to the ith datum and is given by equation (10). Each group $\mathcal{K}_i$ contains its own datum i. We define $\operatorname{rank}(i, \mathcal{K}_i)$ to be the rank of datum i in the set $\mathcal{K}_i$ where the latter is ordered based on the data's index. We have that $\mathtt{C}_{\operatorname{rank}(i,\mathcal{K}_i),|\mathcal{K}_i|}\mathbf{e}_{(\mathcal{K}_i)}$ gives the ith datum's prediction residuals. The Local-Exclusion PRESS statistic and formula are then obtained as:

$$P^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathtt{C}_{\operatorname{rank}(i,\mathcal{K}_i),|\mathcal{K}_i|} \mathbf{e}_{(\mathcal{K}_i)} \right\|_2^2 = \frac{1}{n} \left\| \operatorname{stk}\left(\mathtt{C}_{1,n}(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K}_1,\mathcal{K}_1})^{-1}, \dots, \mathtt{C}_{n,n}(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K}_n,\mathcal{K}_n})^{-1} \right) \left(\mathtt{I} - \hat{\mathtt{A}}\right) \mathbf{b} \right\|_2^2, \tag{13}$$

where the last equality was obtained by using the MEXPRESS formula (6), and some rearrangements. This formula holds with or without the one-datum-one-measurement hypothesis. As the neighborhood size τ shrinks, the Local-Exclusion PRESS statistic approaches the PRESS statistic, and for $\tau = 0$ it gives the PRESS statistic exactly. In practice the neighborhoods may be chosen using other criteria. For instance, a neighborhood system can be defined from the edges of the Delaunay triangulation of the source image points in image registration.

3.5 Other Extensions

The MEXPRESS formula may be used to compute other PRESS statistics non-iteratively, which may be inspired by any types of CV. These include Leave-p-Out PRESS and Random-Sampling PRESS, which may be computed non-iteratively, taking the option of coupled measurements into account. Leave-p-Out PRESS is an exhaustive statistic: each measurement is held out in groups of size p, and all possible $\binom{m}{p}$ groups are covered. Therefore, Leave-p-Out PRESS is rarely used as its computation may be extremely expensive for any p larger than a few data. With the MEXPRESS formula, the Leave-p-Out PRESS statistic can be computed efficiently and non-iteratively. This is because MEXPRESS facilitates the computation of Leave-p-Out PRESS via the inversion of $(p \times p)$ matrices, as can be seen from equation (6). Even though the number $\binom{m}{p}$ of such matrix inverses may be high, the computation is kept tractable for higher values of p than in the iterative approach. Random-Sampling PRESS helds random groups of measurements out. It is non-exhaustive and stochastic. With the MEXPRESS formula, the computation speed of Random-Sampling PRESS can be dramatically improved, allowing one to sample more measurement groups, thereby reducing its statistical bias.

4 Estimating Linear Fractional Warps

Image warps are 2D transformations relating correspondences between a source and a target image. Computing an image warp provides a dense registration between the images.

4.1 General Model

A general warp model is a function $\mathcal{W}$ which maps a point $\mathbf{q} \in \mathbb{R}^2$ in the source image to a point $\mathbf{q}' = \mathcal{W}(\mathbf{q}; \mathtt{P})$ in the target image and depends on a parameter set $\mathtt{P}$. A linear fractional warp has three intrinsic dimensions and may be conveniently modeled using homogeneous coordinates. The parameter set representing the warp's behavior is thus held in a matrix $\mathtt{P} = [\mathbf{p}_1\, \mathbf{p}_2\, \mathbf{p}_3] \in \mathbb{R}^{l \times 3}$, where $l \in \mathbb{N}$ is a varying parameter determining the model's complexity. For a general definition of the linear fractional warps, we use a lifting function $\nu : \mathbb{R}^2 \rightarrow \mathbb{R}^l$ which encapsulates many models such as piecewise affine functions, the TPS and the BBS. A linear fractional warp can then be written as:

$$\mathcal{W}(\mathbf{q}; \mathtt{P}) = \frac{1}{\mathbf{p}_3^\top \nu(\mathbf{q})} \begin{bmatrix} \mathbf{p}_1^\top \nu(\mathbf{q}) \\ \mathbf{p}_2^\top \nu(\mathbf{q}) \end{bmatrix}. \tag{14}$$

Because of the division, the linear fractional warp is not linear in Cartesian coordinates. However, it is linear in homogeneous coordinates as:

$$\begin{bmatrix} \mathcal{W}(\mathbf{q}; \mathtt{P}) \\ 1 \end{bmatrix} \propto \mathtt{P}^\top \nu(\mathbf{q}). \tag{15}$$

The linear fractional warp has $3l - 1$ degrees of freedom. Matrix $\mathtt{P}$ has an undefined scale which may be fixed by imposing a constraint $\lambda(\mathtt{P}) = 0$ which will be discussed shortly. We use the TPS to instantiate the general linear fractional warp model. The TPS depends on l control points whose position in the source image are fixed and whose position in the target image represent the warp's unknown parameters contained in matrix $\mathtt{P}$. We called the resulting warp a Deformable-Perspective Warp (DP-Warp) (Bartoli et al., 2010). The lifting function for the TPS is defined by $\nu(\mathbf{q}) = \mathtt{E}^\top \boldsymbol{\ell}_{\mathbf{q}}$. Matrix $\mathtt{E} \in \mathbb{R}^{(l+3)\times l}$ is related to the bending energy matrix and is constructed from the fixed source control points and the vector $\boldsymbol{\ell}_{\mathbf{q}} \in \mathbb{R}^{l+3}$ holds the kernelized coordinates of the source point with respect to the source control points. More precisely, $\boldsymbol{\ell}_{\mathbf{q}}$ holds $d \log(d)$, which represents the TPS' kernel function (Bookstein, 1989), where d is the distance between $\mathbf{q}$

and the source control points. More details on constructing $\boldsymbol{\ell}_{\mathbf{q}}$ can be found in (Bartoli et al., 2010). The DP-Warp exists for $l \geq 3$. The DP-Warp for $l = 3$ is a homography.

4.2 Linear Least Squares Fitting to Point Correspondences

A usual way to estimate a warp from correspondences is by minimizing the sum of squared distances between the transferred source image points $\mathcal{W}(\mathbf{q}_i, \mathtt{P})$ and the corresponding target image points $\mathbf{q}'_i$, for $i = 1, \ldots, n$. A convex approach can be derived using the algebraic error and leads to a homogeneous LLS cost under the normalization constraint $\lambda(\mathtt{P}) = 0$:

$$\min_{\substack{\mathtt{P}\in\mathbb{R}^{l\times 3}\\ \lambda(\mathtt{P})=0}} \sum_{i=1}^{n} \|\mathbf{e}_i\|_2^2 \qquad \text{with} \qquad \mathbf{e}_i \overset{\text{def}}{=} \mathtt{A}_i\mathbf{p} \in \mathbb{R}^g, \tag{16}$$

where $\mathbf{p} \overset{\text{def}}{=} \text{vec}(\mathtt{P}) \in \mathbb{R}^{3l}$ is the column-wise vectorization of $\mathtt{P}$, $g = 2$ and $\mathtt{A}_i \in \mathbb{R}^{2\times 3l}$. The so-called algebraic distance (Hartley and Zisserman, 2003, Chapter 4) compares two vectors of homogeneous coordinates $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ as $d_A^2(\mathbf{u}, \mathbf{v}) = \|\mathtt{S}(\mathbf{u}\times\mathbf{v})\|_2^2$ with $\mathtt{S} \overset{\text{def}}{=} [\mathtt{I}\ \mathbf{0}] \in \mathbb{R}^{2\times 3}$. Using the target image points' homogeneous coordinates $\text{stk}(\mathbf{q}'_i, 1)$ and the warp in homogeneous form $\mathtt{P}^\top \nu(\mathbf{q}) = \mathtt{P}^\top\mathtt{E}^\top\boldsymbol{\ell}_i$ from equation (15), with $\boldsymbol{\ell}_i \overset{\text{def}}{=} \boldsymbol{\ell}_{\mathbf{q}_i}$, this leads to:

$$\mathbf{e}_i = \mathtt{S}\left(\text{stk}(\mathbf{q}'_i, 1) \times \mathtt{P}^\top\mathtt{E}^\top\boldsymbol{\ell}_i\right) \in \mathbb{R}^2. \tag{17}$$

Using $\mathbf{u} \times \mathbf{v} = [\mathbf{u}]_\times \mathbf{v}$ and some simple algebra we arrive at:

$$\mathtt{A}_i = \mathtt{S}\left[\text{stk}(\mathbf{q}'_i, 1)\right]_\times \text{diag}_3\left(\boldsymbol{\ell}_i^\top\mathtt{E}\right), \tag{18}$$

where diag_f constructs a block-diagonal matrix by duplicating its argument f times. The usual choice for the normalization constraint is $\lambda(\mathtt{P}) = \|\mathtt{P}\|_{\mathcal{F}}^2 - 1$. With this normalization the formulation is homogeneous and quadratic. Even if it has an elegant solution via the SVD, it does not lend itself well into computing PRESS statistics non-iteratively. Without loss of generality, we will use an affine normalization constraint and show that this allows us to compute PRESS statistics non-iteratively, despite the cost being homogeneous.

4.3 Using PRESS to Select the Number of Control Points

Using PRESS to select the number of control points for a model requires *(i)* that the PRESS statistics and formulas be applicable to the model, and *(ii)* that a strategy to vary the number of control points and their placement be provided. These requirements are discussed in the next two paragraphs. Requirement *(i)* is not directly satisfied by the formulation of §4.2. This is because the cost in problem (16) is homogeneous whilst all the PRESS statistics and formulas apply to affine fitting problems with regressors *and* responses. By choosing an affine normalization constraint $\lambda(\mathtt{P}) = 0$ we show however that the problem can be rewritten as an affine problem and the PRESS formulas used.

Computing PRESS for a homogeneous cost. The affine normalization constraint we use enforces $\mathbf{p}_3 \in \mathbb{R}^l$, the third column of $\mathtt{P}$, to have unit mean. Interpreting the linear fractional warps as 3D embeddings, this corresponds to scaling the embedded 3D points so that their average depth becomes one. This constraint is written as:

$$\lambda(\mathtt{P}) \;=\; \lambda(\mathbf{p}) \;\overset{\text{def}}{=}\; \boldsymbol{\lambda}^\top \mathbf{p} - l \qquad \text{with} \qquad \boldsymbol{\lambda} \;\overset{\text{def}}{=}\; \operatorname{stk}(\mathbf{0}, \mathbf{0}, \mathbf{1}) \in \mathbb{R}^{3l}. \tag{19}$$

We can then reparameterize problem (16) using $\mathbf{r} \in \mathbb{R}^{3l-1}$ defined such that:

$$\mathbf{p} \;=\; \mathtt{G}\mathbf{r} - \mathbf{f}, \tag{20}$$

where $\mathbf{f} \in \mathbb{R}^{3l}$ and $\mathtt{G} \in \mathbb{R}^{3l \times (3l-1)}$ are chosen such that $\lambda(\mathtt{G}\mathbf{r} - \mathbf{f}) = 0$, $\forall \mathbf{r} \in \mathbb{R}^{3l-1}$. We may choose $\mathbf{f}$ as any solution of $\lambda(-\mathbf{f}) = 0$ and $\mathtt{G}$ as any basis of $\ker\left(\boldsymbol{\lambda}^\top\right)$. We simply use $\mathbf{f} = -\boldsymbol{\lambda}$ and $G_{i,i} = 1$ for $i \in [1, 3l-1]$, $G_{i+1,i} = -1$ for $i \in [2l+1, 3l-1]$ and $G_{i,j} = 0$ otherwise. The reparameterized problem (16) is:

$$\min_{\mathbf{r} \in \mathbb{R}^{3l-1}} \sum_{i=1}^{n} \|\mathbf{e}_i\|_2^2 \qquad \text{with} \qquad \mathbf{e}_i \;=\; \mathtt{A}_i \mathtt{G}\mathbf{r} - \mathtt{A}_i \mathbf{f}. \tag{21}$$

The PRESS statistic and the new variants we propose can then be computed non-iteratively using the corresponding PRESS formulas. All computations are done in image coordinates normalized to $[-1, 1]^2$. Because the algebraic error is defined up to an arbitrary scale factor inherited from matrix $\mathtt{P}$, we consistently rescale it so that its average magnitude lies around a few hundred units.

Selecting the number of control points. We automatically select the number l of control points in linear fractional warps by minimizing PRESS. The idea is to simply compute the chosen PRESS statistic $P^2(l)$ for $l = l_{\min}, \ldots, l_{\max}$ and keep the l with the lowest PRESS statistic. We use $l_{\min} = 3$ and set $l_{\max}$ according to the number of correspondences.[1] Cross-Validation (and thus the PRESS) have been successfully used to prevent overfitting in model training followed by testing. The analogy with an image warp is that the training stage corresponds to fitting the warp to given sparse correspondences and the testing stage corresponds to evaluating the warp at points off these correspondences. We place the control points as uniformly as possible using Lloyd's algorithm (Lloyd, 1982). We measure all main types of PRESS

[1] We use $l_{\max} = 6$ for $n = 10$, $l_{\max} = 7$ for $n = 15$ and $l_{\max} = \min(\operatorname{round}(\frac{m}{2}), 100)$ otherwise.

mentioned in §3. Recall that in Local-Exclusion PRESS, each point is used as validation set in turn as in regular PRESS. The difference is that using MEXPRESS, subsets of data can be held out simultaneously with each point. We use two main strategies to define the held-out subsets:

- *The Exclusion Radius strategy.* We exclude all points within r pixels of the source image test point, where r is defined as a fraction of the image diagonal.
- *The Nearest Neighbors strategy.* We exclude the k closest points to the source image test point. We choose k as a fixed number or as a fraction of the number of correspondences n.

These two strategies behave very differently with respect to the density of point correspondences and control points.

4.4 Experimental Results

The number n of correspondences is the most important parameter when trying to fit a warp. We want to observe the behavior of the various PRESS statistics with non-iterative formulas in function of n. Our goal is not to establish a definitive criterion to choose a warp's best number of control points. This is a very difficult problem, in particular because no ground truth can be easily defined. Our goal is to show that with MEXPRESS, many PRESS statistics can be computed non-iteratively and may offer new possibilities to study this problem.

4.4.1 Compared PRESS Statistics

We monitor the fitting RMSR FIT and the following PRESS statistics, computed using their non-iterative formulas:

- PRESS. The basic PRESS statistic (3), which does not take coupled measurements into account, and therefore almost always underestimates Coupled-Measurement PRESS.
- CMPRESS. The Coupled-Measurement PRESS statistic (9), §3.2.
- KF10PRESS and KF05PRESS. The k-Fold PRESS statistic (9) with $k = 10$ and $k = 5$ folds, §3.3.
- ER01PRESS, ER05PRESS and ER10PRESS. The Local-Exclusion PRESS statistic (13) with an Exclusion Radius of 1%, 5% and 10% of the image diagonal.
- NN02PTSPRESS, NN05PTSPRESS and NN10PTSPRESS. The Local-Exclusion PRESS statistic (13) excluding the 2, 5 and 10 Nearest-Neighbors.

- NN01PCTPRESS, NN05PCTPRESS and NN10PCTPRESS. The Local-Exclusion PRESS statistic (13) excluding the x Nearest-Neighbors, where x corresponds to 1%, 5% and 10% of the number of correspondences, respectively.

Except PRESS, all these statistics take measurement coupling into account, with $g = 2$. We also measured a test error, which we optimize with respect to the number of control points. For simulated data, this is called TEST and defined as the average root mean squared transfer error $\frac{1}{|\Omega|}\sqrt{\int_\Omega d_A^2(\mathcal{W}(\mathbf{q};\mathtt{P}),\mathbf{q}')\ \mathrm{d}\mathbf{q}}$, where $\Omega \subset \mathbb{R}^2$ is the source image domain. For real data, this is a Jaccard index, computed as follow. We mark the surface's contour in the source and target images prior to warp estimation. We then use the estimated warp to transfer the source surface's contour to the target image and compute the Jaccard index with the target surface's contour. The optimal number of control points is selected which maximizes (and not minimizes) the Jaccard index.

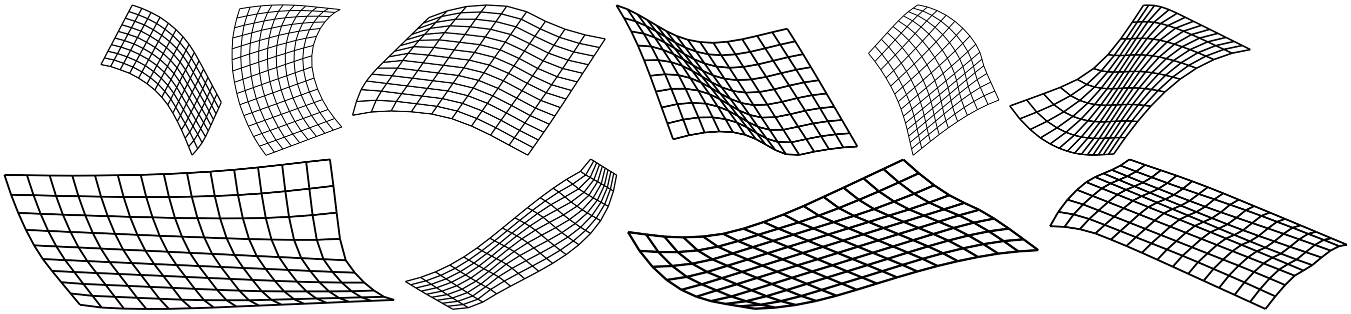


Figure 1: The batch of 10 simulated surfaces seen from some of the randomly generated camera poses.

4.4.2 Simulated Data

We simulated a deformable surface using a rescaled isometric model (Perriollat and Bartoli, 2013) in 10 different configurations shown in figure 1. These surfaces were created by a 3D embedding of the 2D plane, which we also used to create random 3D point correspondences. We then simulated an HD camera with a 35 mm sensor and a variable focal length in the range $[20, 50]$ mm, observing each surface from various poses. Image pairs were formed by using the 90 possible pairs of surfaces and 5 random poses and focal length, yielding a total of 450 possible configurations. A centred Gaussian noise with 1 px standard deviation was added to all simulated image points.

We varied the number n of point correspondences in $\{15, 30, 150, 500\}$. In warp estimation, $n = 15$ represents very few correspondences, $n \in \{30, 150\}$ are typical numbers of correspondences and $n = 500$ represents many correspondences. Increasing the number of correspondences allows one to detect the criteria which tend to overfit as they unreasonably increase the number of control points. The results are given in

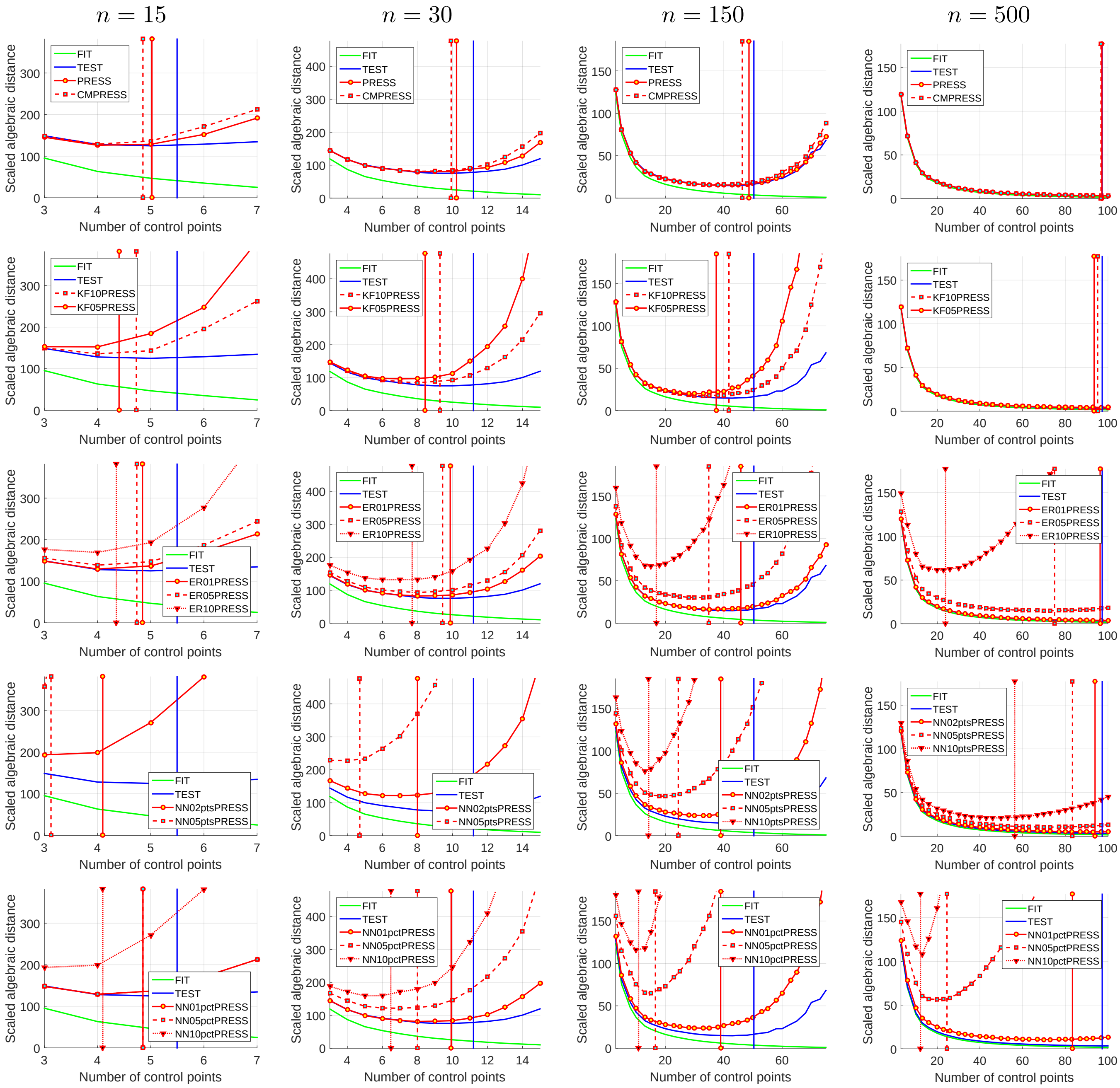


Figure 2: Behavior of the PRESS statistics, fitting residual FIT and test error TEST as a function of the warp's number of control points (horizontal axis of each graph) and number of point correspondences (columns of the figure). The vertical lines represent the average number of control points selected by the various PRESS statistics and TEST.

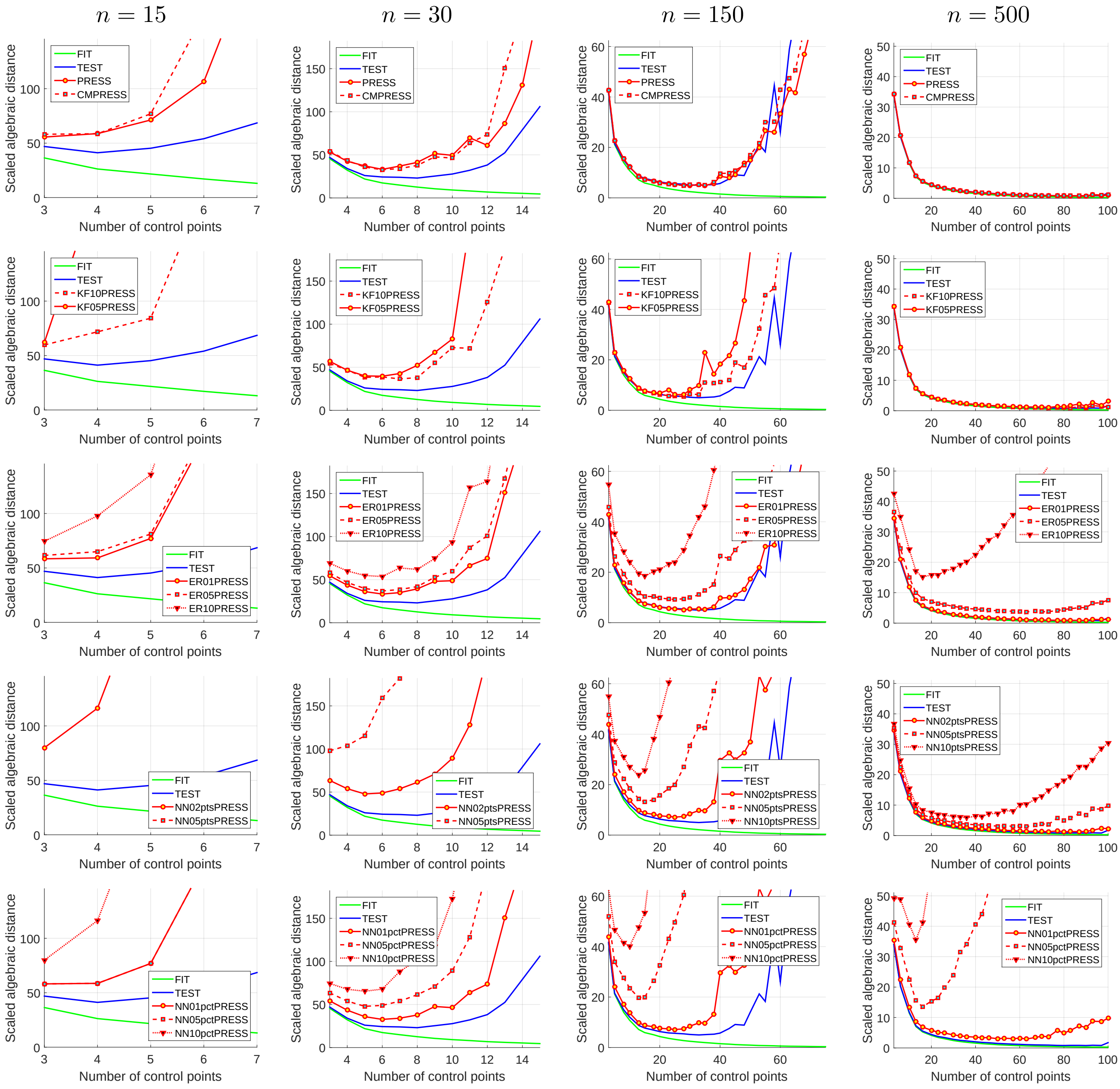


Figure 3: Standard deviation of the PRESS statistics, fitting residual FIT and test error TEST as a function of the warp's number of control points (horizontal axis of each graph) and number of point correspondences (columns of the figure).

	$n = 15$		$n = 30$		$n = 150$		$n = 500$	
Criterion	l	FIT	l	FIT	l	FIT	l	FIT
TEST	05.5±01.3	42.5±28.4	11.2±02.2	21.8±11.9	50.3±08.9	04.0±02.0	97.3±03.8	01.7±00.3
PRESS	05.0±01.3	42.4±26.4	10.2±02.6	23.3±15.2	48.6±09.6	04.0±02.1	97.1±03.8	01.7±00.3
CMPRESS	04.9±01.3	45.2±28.9	09.9±02.6	24.7±15.3	46.3±09.4	04.4±02.3	96.7±04.4	01.7±00.3
KF10PRESS	04.7±01.3	47.7±29.7	09.3±02.4	27.7±16.1	41.8±07.8	05.4±02.4	95.2±05.6	01.8±00.4
KF05PRESS	04.4±01.1	53.1±31.3	08.4±02.1	31.8±16.6	37.4±06.9	06.6±02.9	93.5±06.6	01.9±00.4
ER01PRESS	04.8±01.3	45.4±29.0	09.9±02.6	25.0±15.7	45.9±09.5	04.5±02.3	96.3±04.6	01.7±00.3
ER05PRESS	04.7±01.3	47.0±29.5	09.4±02.5	26.8±16.5	34.9±08.1	07.5±03.7	75.0±12.8	02.7±00.9
ER10PRESS	04.4±01.2	52.8±31.3	07.7±02.5	36.9±22.6	16.9±04.5	19.6±07.5	23.9±07.7	14.9±06.9
NN02PTSPRESS	04.1±01.1	58.7±34.3	08.0±02.5	35.1±21.0	38.9±08.5	06.2±03.1	93.9±06.8	01.8±00.4
NN05PTSPRESS	03.1±00.4	88.9±38.0	04.7±01.6	67.3±34.4	24.4±07.1	13.0±06.1	83.4±11.9	02.3±00.8
NN10PTSPRESS	—	—	—	—	14.3±03.5	23.6±08.5	56.3±12.1	04.5±01.9
NN01PCTPRESS	04.9±01.3	45.2±28.9	09.9±02.6	24.7±15.3	38.9±08.5	06.2±03.1	83.4±11.9	02.3±00.8
NN05PCTPRESS	04.9±01.3	45.2±28.9	08.0±02.5	35.1±21.0	16.6±04.7	20.3±08.1	24.6±07.1	14.4±06.5
NN10PCTPRESS	04.1±01.1	58.7±34.3	06.5±02.3	47.1±27.4	10.8±03.3	34.0±17.9	12.2±03.4	31.2±15.4

Table 1: Average and standard deviation of the selected number of control points l and fitting RMSR FIT for the different criteria for $n \in \{15, 30, 150, 500\}$ point correspondences.

figure 2. We observe that the fitting RMSR decreases steadily as the number of control points increases, for all numbers of correspondences. In other words, we do not observe a clear transition point or zone, which would allow us to figure out a reference number of control points, against which to evaluate the different criteria in terms of underfitting and overfitting. We thus evaluated the criteria differently. We used the fact that, given that the data model is the same set of surfaces for all numbers of correspondences, changing the number of correspondences should have a limited impact on the number of selected control points. We do not expect however that the number of selected control points does not change at all, as we varied the number of correspondences between 15 and 500, which represent extreme possibilities. In order to understand the relationship between the number of selected control points and the number of correspondences, we computed summary statistics, namely the average and standard deviation of the selected number of control points, over the 450 simulated geometric configurations. These statistics are given in table 1 for the four numbers of simulated correspondences. The average number of control points is also visible in figure 2. We made the following observations:

- PRESS and CMPRESS (first row of figure 2) have a very similar behavior. Because of its larger bias in prediction accuracy, the former favors slightly larger numbers of control points. Both select numbers of control points very close to TEST. They do fine for small numbers of correspondences, but quickly overfit as the number of correspondences grow, as the number of selected control points increases from approximately 5 to almost 100. This is because in these criteria, validation is computed on data very similar to the training data when the data density increases, and models with higher complexity are then not penalized enough.

- KF10PRESS and KF05PRESS (second row of figure 2) mitigate overfitting compared to CMPRESS, but end up overfitting too when the number of correspondences grow large. The number of selected control points increases from fewer than 5 to approximately 95. This is because in these criteria, the folds are selected randomly. Some of the held out correspondences may thus be in the vicinity of the validation correspondences, and thereby penalize complexity to a larger extent. Despite this advantageous property, k-Fold PRESS is stochastic and non-exhaustive, which is not a desirable property.

- ER01PRESS, ER05PRESS and ER10PRESS (third row of figure 2) have different behaviors, showing that the exclusion radius plays an important role. For all three criteria the number of selected control points is lower than, but close to, 5 for 15 point correspondences. It however changes very differently when the number of point correspondences increases. With a small exclusion radius of 1% of the image diagonal, the behavior is very similar to CMPRESS. The number of point correspondences increases to almost 100. This is understandable as very few correspondences are excluded from training, and higher complexity is thus not penalized much. With a larger exclusion radius of 5% of the image diagonal, we clearly observe a reduction of overfitting for large densities of correspondences, as the number of selected control points is then 75. As expected, this effect is amplified by increasing the exclusion radius to 10% of the image diagonal, and overfitting is substantially limited, even for the larger number of 500 correspondences, as the number of selected control points is then lower than 25.

- NN02PTSPRESS, NN05PTSPRESS and NN10PTSPRESS (fourth row of figure 2) do not mitigate overfitting as efficiently as the exclusion radius. This is because they use a fixed number of correspondences in the exclusion sets, which represents a very tiny fraction of the correspondences when the correspondence density grows larger. For smaller numbers of correspondences, this naturally has the opposite effect, and causes significant underfitting. The number of selected control points varies between approximately 4 and 94 for NN02PTSPRESS and between approximately 3 and 83 for NN05PTSPRESS. NN10PTSPRESS cannot be computed for settings with 15 and 30 correspondences, and spans between approximately 14 and 56 selected control points for 150 and 500 point correspondences respectively.

- NN01PCTPRESS, NN05PCTPRESS and NN10PCTPRESS (fifth row of figure 2) have a similar behavior to the statistics using the exclusion radius, as they largely mitigate overfitting. This is because using a fraction of the number of correspondences in the exclusion sets (as opposed to a fixed number) naturally adapts to the number of correspondences. Their selected number of control points varies between almost, but fewer than, 5 for all three criteria, and 83, 25 and 12 respectively.

The graphs in figure 2 were obtained by averaging measurements obtained over 450 geometric configurations. The corresponding standard deviation is given in figure 3. We observe that it decreases for larger numbers of correspondences and increases for larger numbers of control points. More precisely, it is sufficiently small relatively to the observed average for $n = 150$ and $n = 500$ correspondences, making our observations perfectly valid. For $n = 15$ and $n = 30$ however, it is too large to conclude that our general observations consistently hold. Our goal however is not to analyze the absolute value of the criteria on their own, but rather to understand if they lead to a sensible and stable selection of the number of control points, in terms of how they mitigate underfitting and overfitting. For this reason, we use the summary statistics given in table 1. We observe that the fitting RMSR may have a large standard deviation for smaller numbers of correspondences, for all criteria. This is also the case for larger numbers of correspondences for criteria which involve holding out a larger number of correspondences in their internal training phase, namely NN10PCTPRESS and ER10PRESS. These observations were expected from the standard deviations already observed in figure 3. However, and this is the important observation, for all numbers of correspondences and all criteria, the selected number of control points has a reasonably small standard deviation. This makes our conclusions regarding the capacity of each criterion to mitigate underfitting and overfitting perfectly valid. This also suggests that, while the fitting RMSR and the value of the criteria may vary significantly depending on the location of the correspondences, the selected number of control points is eventually quite independent of this location and stable.

4.4.3 Real Data

We used images of a poster taken by a digital camcorder with a 4.23 × 3.17 mm sensor and a variable focal length in the range $[4.60, 24.48]$ mm. The image size was 816 × 612 px and the focal length range translated to $[888, 4724]$ px. 206 point correspondences were manually labelled on all images. We combined various scene geometries and imaging conditions, rigid-flat/rigid-non-flat/deformable scene with perspective/affine imaging, and sampled $n \in \{51, 103, 206\}$ correspondences in each configuration. Following our observations on simulated data, we excluded the k-Fold PRESS (KF10PRESS and KF05PRESS) and Local-Exclusion with a fixed number of points PRESS (NN02PTSPRESS, NN05PTSPRESS and NN10PTSPRESS) statistics from the results, as these strategies do not mitigate overfitting well and sometimes even underfit.

Rigid-flat scene (figure 4 and table 2). The flat scene is the only case for which the true number of control points is known and equal to three. This is the simplest case, whose geometry can be exactly explained by the DP-Warp, for both the affine and the perspective imaging conditions. We observed that all criteria lead to overfitting in some conditions. This includes JACCARD INDEX, despite its use of the surface's contour as supplementary 'test' data. Though being an independent test criterion, it is therefore clearly unreliable. We observe that overfitting increases with the number of correspondences and from affine to perspective conditions. This is understandable as more data tends to better constrain more complex models, and perspective conditions require a more advanced model than affine conditions. We observe that PRESS, CMPRESS and ER01PRESS give the same results and are largely prone to overfitting. They are very closely followed by NN01PCTPRESS, which exhibits the same behaviour. By tightening the exclusion strategy, which means by excluding more correspondences, Local-Exclusion PRESS improves its efficiency at limiting overfitting. This is observed for, in order of increasing efficiency, the following criteria, ER05PRESS, ER10PRESS, NN05PCTPRESS and NN10PCTPRESS.

	Affine imaging			Perspective imaging		
Criterion	$n = 51$	$n = 103$	$n = 206$	$n = 51$	$n = 103$	$n = 206$
JACCARD INDEX	3	3	3	5	22	16
PRESS	4	10	16	7	10	33
CMPRESS	4	10	16	7	10	33
ER01PRESS	4	10	16	7	10	33
ER05PRESS	4	6	6	7	10	10
ER10PRESS	4	5	6	7	10	10
NN01PCTPRESS	4	10	10	7	10	20
NN05PCTPRESS	4	3	6	4	10	6
NN10PCTPRESS	3	3	6	3	5	6

Table 2: Selected number of control points l for different criteria and the cases shown in figure 4. Because the surface is flat in both images, the 'true' number of control points is three.

Rigid-non-flat and deformable scenes (figures 5 and 6). In these two cases it is difficult to deem that a statistic does better than another, as ground-truth cannot be known. We can however make reliable qualitative observations. In all four cases we have that PRESS and CMPRESS overfit. All methods based on the Local-Exclusion strategy mitigate overfitting efficiently to a degree consistently related to the strength of the correspondence exclusion rule. Both strategies (the exclusion radius and the percentage of nearest-neighbors) do equally well. For stronger exclusion rules, the statistics may even underfit, as for instance NN10PCTPRESS in the rigid affine case with 103 correspondences.

4.4.4 Summary and Discussion

On the one hand, the usual PRESS and coupled measurements corrected CMPRESS statistics almost always overfit, for both simulated and real data. On the other hand, the PRESS statistics using Local-Exclusion have very interesting behaviors. In particular, the strategies of using an exclusion radius and a fraction of the number of correspondences nearest-neighbors to form the exclusion sets seem to resolve overfitting for large numbers of correspondences, without causing underfitting for smaller numbers of correspondences.

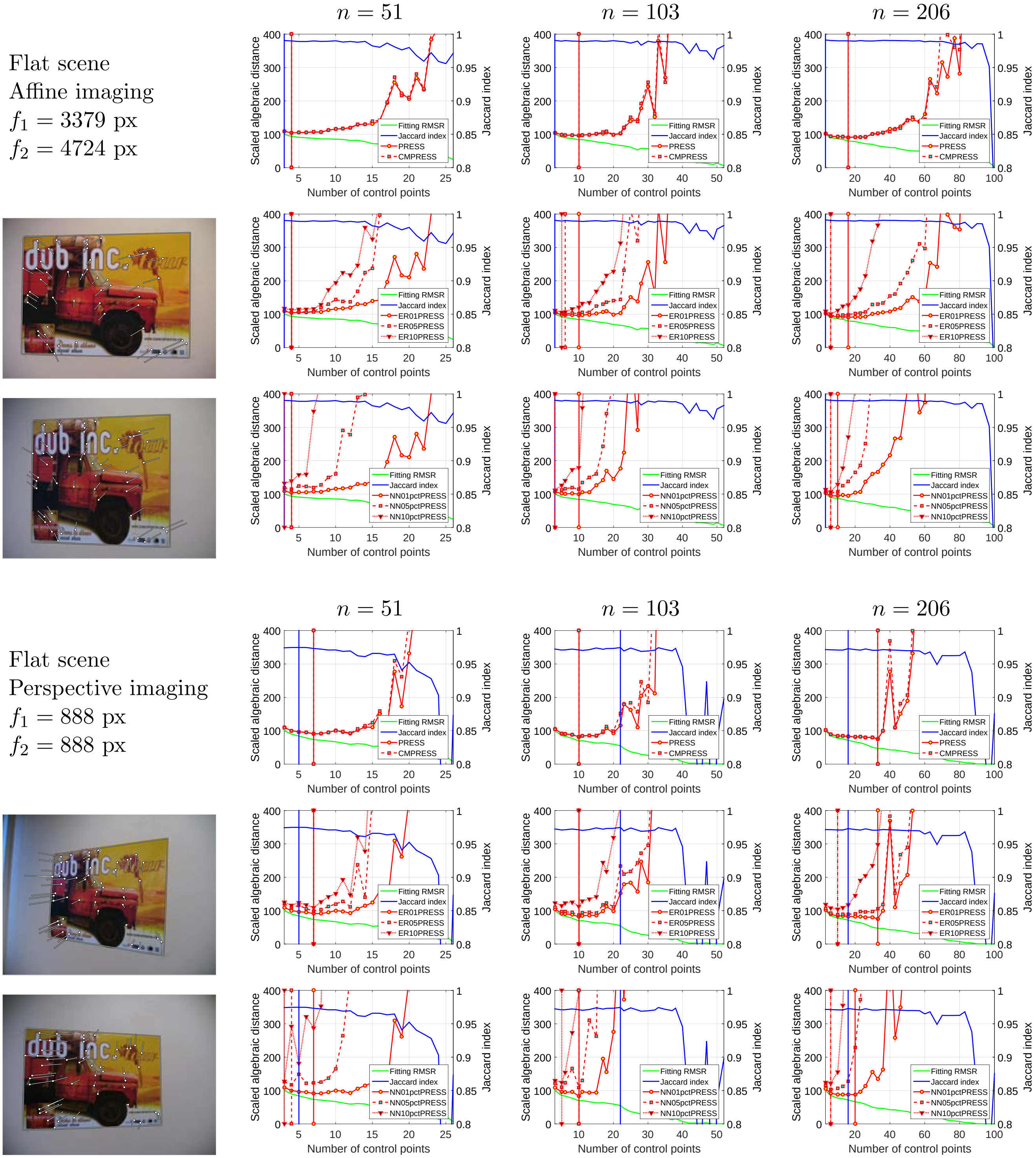


Figure 4: Real data, rigid-flat cases. Behaviour of the PRESS statistics, fitting residual FITTING RMSR, Jaccard index between the target and registered source surface's contours JACCARD INDEX as a function of the warp's number of control points (horizontal axis of each graph) and number of correspondences (columns of the figure). The vertical lines represent the number of selected control points for each criterion, which are also given in table 2. Because the surface is flat in both images, the 'true' number of control points is three. The source and target images overlaid with the correspondences and surface's contour are shown on the left.

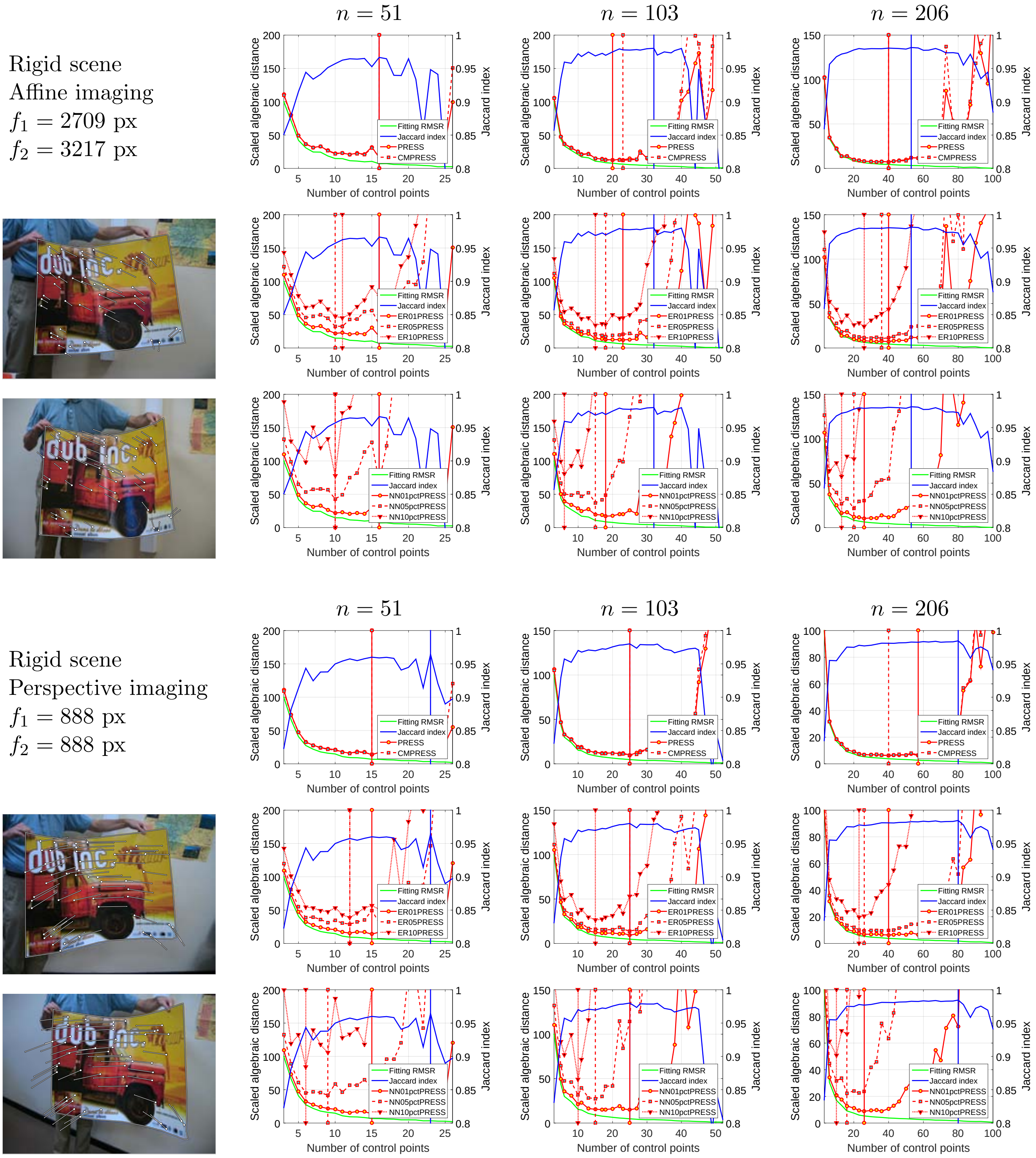


Figure 5: Real data, rigid-non-flat cases. Behaviour of the PRESS statistics, fitting residual FITTING RMSR, Jaccard index between the target and registered source surface's contours JACCARD INDEX as a function of the warp's number of control points (horizontal axis of each graph) and number of correspondences (columns of the figure). The vertical lines represent the number of selected control points for each criterion. The source and target images overlaid with the correspondences and surface's contour are shown on the left.

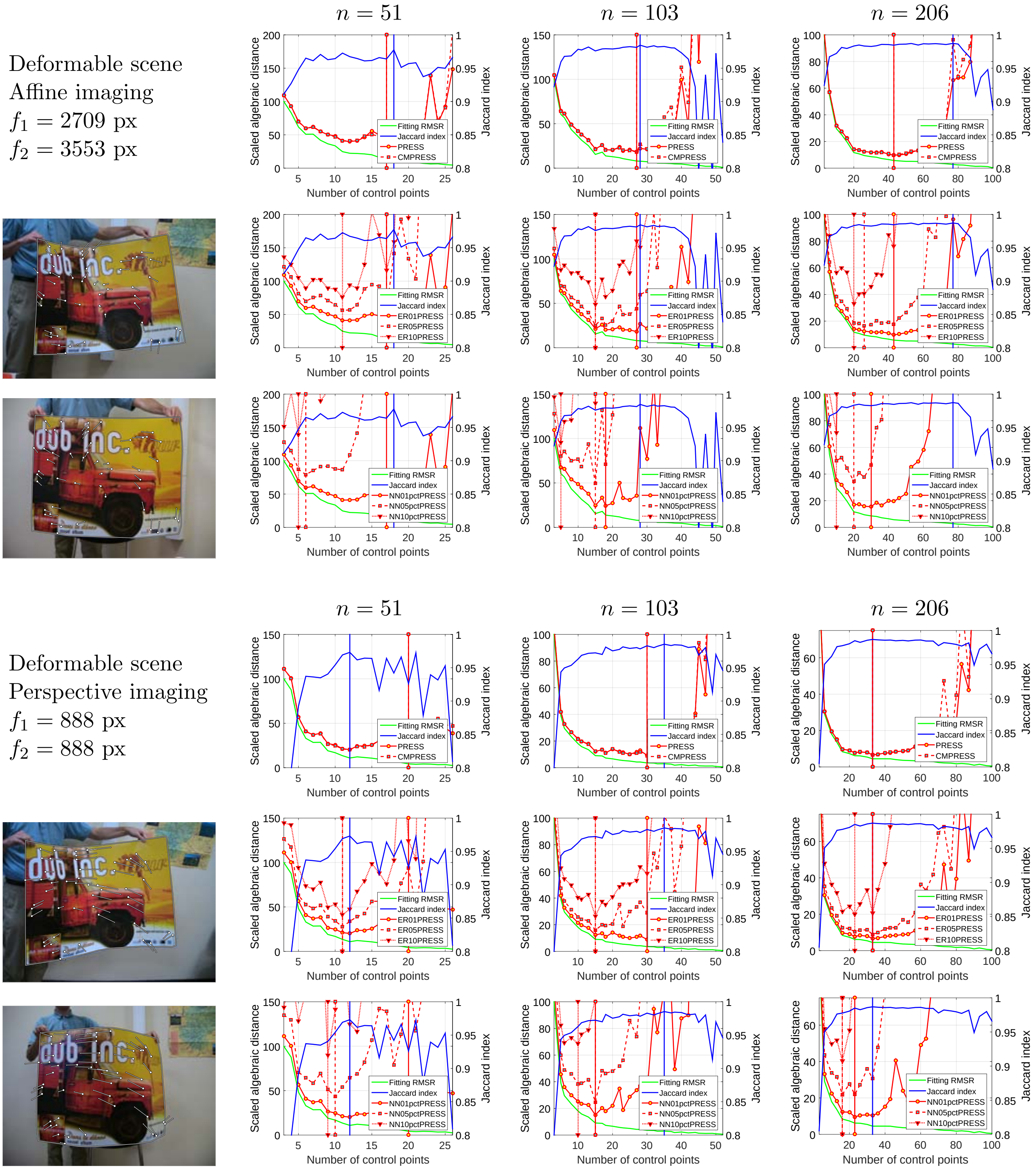


Figure 6: Real data, deformable cases. Behaviour of the PRESS statistics, fitting residual FITTING RMSR, Jaccard index between the target and registered source surface's contours JACCARD INDEX as a function of the warp's number of control points (horizontal axis of each graph) and number of correspondences (columns of the figure). The vertical lines represent the number of selected control points for each criterion. The source and target images overlaid with the correspondences and surface's contour are shown on the left.

The exclusion radius strategy is advantageous, as using a fixed number of nearest-neighbors is more prone to underfitting and overfitting with smaller and larger numbers of correspondences, respectively. The exclusion radius strategy has two more advantages. First, its parameter corresponds to the image area that an image warp should be able to fill in from neighboring correspondences. From empirical observations, we recommend choosing between $5-10\%$ of the image diagonal. Second, it naturally adapts to the correspondence density, when the correspondences are not uniformly spread in the image. We observed that for smaller numbers of correspondences, the criteria holding out larger numbers of correspondences in their internal training phase, namely the Local-Exclusion PRESS with a large exclusion radius or a large percentage of nearest-neighbors, may have an unstable behavior when the number of control points grows larger. This was also revealed by the standard deviation shown in figure 3. This is explained by the fact that in these cases, these warps have difficulties predicting the held out correspondences. This both increases the value of the criterion, and thus prevents overfitting as desired, and its standard deviation. In almost all cases, these instabilities occurred for numbers of control points larger than the number selected according to the criterion. In other words, the instabilities did not disturb the process of selecting the number of control points. This was confirmed by the low standard deviation of the selected numbers of control points for these criteria shown in table 1.

Handling coupled measurements in CMPRESS does not make a strong difference with PRESS in all cases. We noticed that this difference increases with the compound effect of smaller numbers of correspondences and larger numbers of control points. This is because with fewer data and a more flexible model, prediction becomes more difficult, and the effect of using all measurements corresponding to a datum to predict one of the measurements thus becomes stronger. Not handling coupled measurements thus artificially improves the prediction accuracy, and causes PRESS to underestimate CMPRESS. However, we did not notice that this discrepancy produced a significant difference in the selected number of control points in our experiments. This does not mean that for other types of problems, especially with higher dimensions, this difference may also be negligible.

The selected number of control points is given by finding the minimum PRESS value. In some cases however, this minimum value may be ill-defined, for two reasons. The first reason is that there may be multiple minima with similar values. In figure 4 for instance, in the perspective case and for $n = 103$ correspondences, ER10PRESS overfits by selecting 10 control points, but a local minimum with a very similar value exists for 5 control points, and lies closer to 3, the true number of control points. The second reason is that the minimum may lie in a shallow part of the PRESS curve. In figure 4 for instance, in the perspective case and for $n = 206$ correspondences, NN01PCTPRESS overfits by selecting 20 control points, while a range of similar values exists between 10 and 20 control points. We have not found that this phenomenon of an ill-defined minimum PRESS value was specific to a criterion, experiment or setup.

5 Fitting Surfaces

We show that MEXPRESS can be used to select the number of control points used in a continuous surface function model. The control points are chosen iteratively to minimize the reconstruction residual, and two application cases are considered, with dense and sparse data, obtained by a depth sensor and SfT, respectively.

5.1 General Method

We follow the same general method and algorithm for both application cases.

5.1.1 General Points

In curve and surface fitting, when using a function model with a linear basis, one has two key ingredients to choose: *(i)* the control point placement strategy and *(ii)* a criterion which allows one to select the number of control points. Our strategy is to place the control points iteratively where the fitting residual is highest, monitor the value of the selection criterion, and choose the number of control points for which this selection criterion has the smallest value.

5.1.2 Iterative Control Point Placement

Control Point placement strategies fall into three main categories. In the first category, the control points are added iteratively based on the per-datum residual error. They are typically positioned at those places where the fitting is poor. This strategy was followed for instance for fitting a least squares BBS to scatter data (Dierckx, 1981) and to mesh data (Dierckx, 1993). In the second category, the control points are added and pruned from a prespecified discrete set of locations (Breiman, 1993; Friedman, 1991). In the third category, the control points are optimized jointly with the function's parameter set. This strategy was used in a number of approaches, which fix the number of control points a priori (Süssmuth et al., 2010), iteratively grow the set of control points (Molinari et al., 2004) or jointly optimize the number of control points (Miyata and Shen, 2005). The first and second categories use convex optimization. The third category involves a nonconvex cost prone to numerous local minima (Jupp, 1978), solved for instance with Levenberg-Marquardt (Süssmuth et al., 2010), non-deterministic sampling (Molinari et al., 2004) and evolutionary computing (Miyata and Shen, 2005). Our approach falls in the first category. We start with the minimum number of $l = 3$ control points, for which the fitted surface is a plane, and iteratively add control points. At each iteration, a candidate control point is found by combining three constraints. First, the control point must lie within the function's domain $\Omega \subset \mathbb{R}^2$ defined by the input data. Second, the control point must lie at a minimum, predefined distance $\kappa \in \mathbb{R}$, to all the other control points. This is to favor an even spreading of the control points, and to avoid the 'lethargy' problem, causing control points to sometimes accumulate at the same position (Jupp, 1978). Third, the control point is chosen where the fitting is the poorest, to reduce the fitting residual as best as possible. Because Ω is a compact subset, choosing $\kappa > 0$ implies that the algorithm necessarily terminates in a finite number of iterations. However, our goal is to use a selection criterion to limit the number of used control points and thus the model's complexity.

5.1.3 Selection Criteria

The selection criterion is an essential component as it determines the number of control points to be used for a dataset. A first type of strategy is to have the user to provide additional knowledge on the expected fit, such as the tradeoff between the fitting residual and the surface's smoothness (Dierckx, 1981) or an upper bound on the fitting residual (Dierckx, 1993). A second type of strategy is to use a criterion inspired from model selection such as Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978), both used in (Miyata and Shen, 2005; Molinari et al., 2004), and k-Fold Cross-Validation, used in (Breiman, 1993). We propose to use the PRESS statistics, and in particular the Local-Exclusion PRESS, as selection criteria. This is because we do not expect the user to provide additional knowledge on the expected fit, and want to be able to handle datasets with various types of density.

5.1.4 Algorithm

The following algorithm implements the above-described iterative control point placement and stopping criteria:

1. Initialize the control point set $\mathcal{C}$ to three random points in Ω separated by a distance greater than κ
2. Fit the model with the control points in $\mathcal{C}$, estimate and store the selection criterion $\xi(\mathcal{C})$
3. Estimate the per-datum fitting residual
4. Attempt to draw a new control point $\mathbf{c} \in \Omega$ separated by a distance greater than κ to all control points in $\mathcal{C}$ and which maximizes the fitting residual
5. If $\mathbf{c}$ exists then update $\mathcal{C} \leftarrow \mathcal{C} \cup \mathbf{c}$ and loop to step 2
6. Return $\mathcal{C}$ such that $\xi(\mathcal{C})$ is minimized

In practice, we have implemented this algorithm as PRESS using the PRESS statistic (3), as CMPRESS using the Coupled-Measurement PRESS statistic (9), §3.2, as ER01PRESS, ER05PRESS, ER10PRESS using the Local-Exclusion PRESS statistic (13) with an Exclusion Radius of 1%, 5% and 10% of the domain Ω's diagonal length and as NN01PCTPRESS, NN05PCTPRESS, NN10PCTPRESS using the Local-Exclusion PRESS statistic (13) excluding a number of Nearest-Neighbor data points given by respectively 1%, 5% and 10% of the total number of data points. We have also implemented the BIC and AIC criteria as BIC and AIC. These two criteria have numerical values which have a very different scale (they are usually much larger) compared to the PRESS statistics and FIT. However, what matters is their minimizers. Therefore, in order to display them on the same graphs as the other criteria, we positively rescaled and offset them so that they align to FIT as best as possible in the least-squares sense. This does not change their minimizers and preserve their graph's shapes.

5.2 Dense Depth-Maps from Depth Sensor

We applied the proposed general fitting algorithm to fit a surface to depth-maps extrated from RGBD images captured by Kinect v2. More precisely, we fit a TPS to interpolate the depth channel. This means that the target space has dimension one, $g = 1$. Therefore, the PRESS and Coupled-Measurement PRESS statistics are equivalent and we only give the former. The depth-map data are extremely dense, and we subsampled them to 10,000 points. We may expect that some selection criteria overfit. Our work is complementary to most works in RGBD image processing, which address the problems of surface completion and registration between multiple views (Steinbrücker et al., 2013; Xu et al., 2014; Zollhöfer et al., 2014). We show results on two datasets, for which we limited the maximum number of control points to 175.

The first dataset shows a smooth surface of size 30 × 45 cm. The input image and results are shown in figure 7. We observe that FIT transitions between a steep and a mild decrease at approximately 27 control points, for a fitting RMSR of about 2 mm. This makes this dataset interesting, for 27 is then the number of control points that one would manually choose, and may serve as a reference to evaluate the tested criteria. In this respect, we observe that NN10PCTPRESS largely underfits, by selecting only 7 control points, that BIC, ER01PRESS, ER05PRESS and PRESS overfit, by selecting between 64 and 81 control points, and that AIC largely overfits, by selecting 171 control points. However, ER10PRESS and NN05PCTPRESS select 27 and 28 control points respectively, matching the manually established reference. We observe that the reconstructed surfaces do not exhibit strong differences. This is because the dataset is dense and has limited measurement noise, and because the surface is smooth, matching the estimated smooth TPS model, restricting the modeling error to a large extent. Nonetheless, we observe that by underfitting NN10PCTPRESS increases the fitting RMSR of the reference solution by a factor of approximately 3.5, while by overfitting BIC, ER01PRESS, ER05PRESS and PRESS decrease it only by a factor of approximately 0.7. We observe that the control points are initially placed at locations where the surface has its highest bending. Due to the minimal inter-control-point distance criterion, they however finally cover the complete surface.

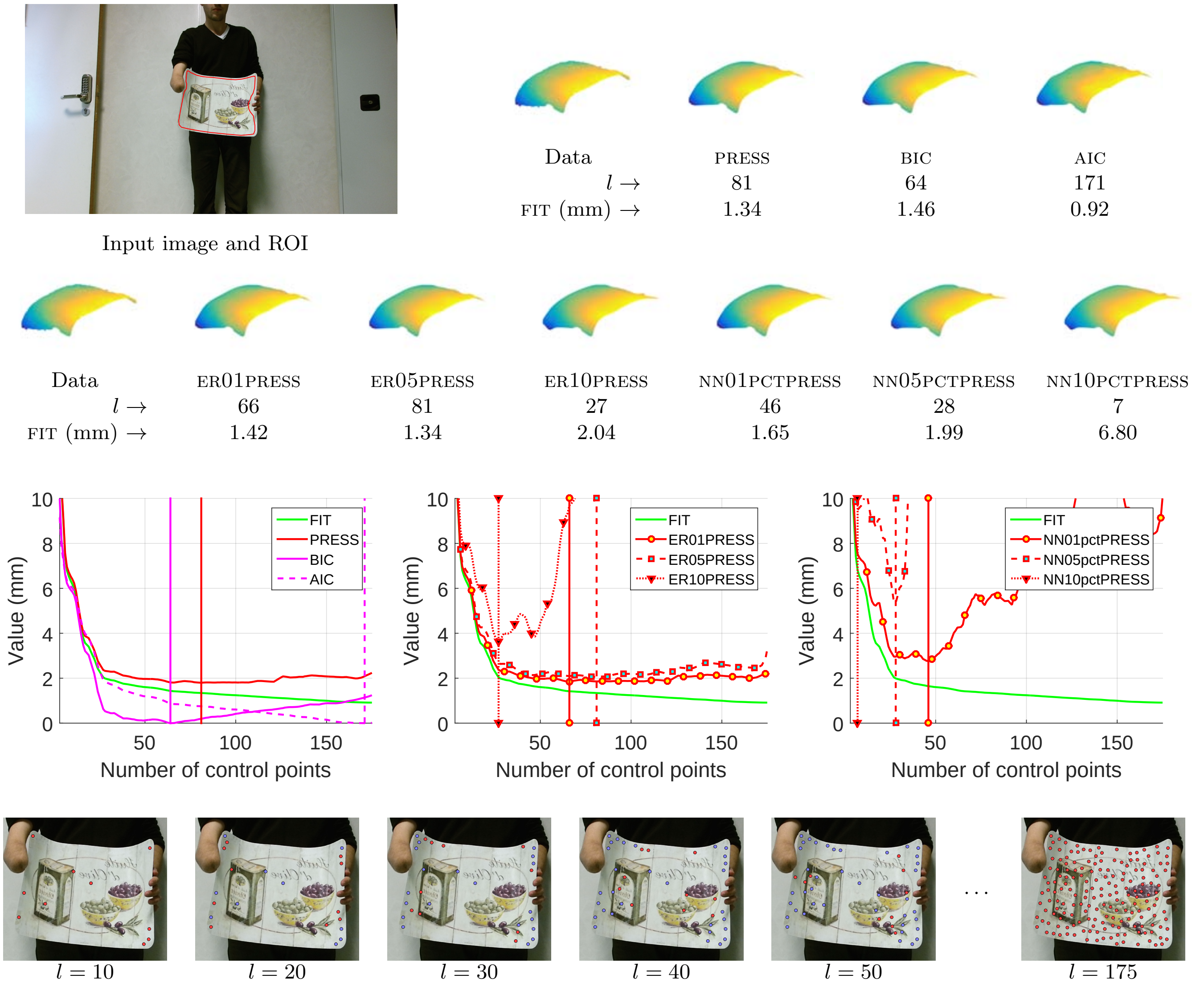


Figure 7: **Surface fitting for dense depth-maps: the smooth dataset.** The first two rows show the input image and the fitted surface for the different selection criteria. The third row shows the various selection criteria as a function of the number of control points. The y-axis of the graphs, labelled 'value', is in mm for all criteria as they give a value in depth units, except AIC and BIC. The fourth row shows the automatically placed control points in blue, with the last 10 placed points in red. The complete set is shown in red, right-most.

The second dataset shows a stack of four books of width 24 cm and height 18 cm. The surface formed by the binding of these four books has sharp angles. We may thus expect more control points to be necessary to model this dataset than the smooth one. The input image and results are shown in figure 8. We do not observe a clear transition in FIT as in the smooth case. However, we observe clear differences between the reconstructed surfaces and use this as our main evaluation criterion. This is because, though the dataset is dense and has limited measurement noise, the sharp folds of the surface introduce a modeling error for the TPS we estimate is a smooth model. In this respect, we observe that nn05pctpress and nn10pctpress largely underfit, selecting respectively 15 and 3 control points, as they fail to capture the surface's geometry. Note that 3 is the minimum number of control points. On the other hand, er01press, er05press, bic, aic and press overfit, selecting between 162 and 175 control points. Note that 175 is the maximum number of control points, and is selected by aic. Overfitting is observed because the reconstructed surfaces for these criteria capture the surface's geometry approximately well but introduce undesired high frequencies. Finally, er01press and nn01pctpress select respectively 97 and 113 control points. They capture the surface geometry to a similar extent, with a fitting RMSR of 2.91 mm and 2.59 mm, respectively. As expected, overfitting reduces the fitting RMSR by a limited factor, as, for instance, er05press lowers it to 1.88 mm. On the other hand, underfitting increases it to 8.36 mm for nn10pctpress, for instance. We observe that the control points are initially placed at the frontiers between the books, creating sharp folds in the surface.

These experiments suggest that many existing criteria lead to overfitting when applied to dense data. BIC is the existing criterion which best mitigates overfitting. We observed that the proposed PRESS statistic based on an exclusion radius of 10% of the domain does not overfit.

5.3 Sparse Points from Shape-from-Template

SfT is a problem where one wishes to recover the deformation of an object's model to match an input image (Perriollat et al., 2011; Salzmann et al., 2007). The vast majority of methods recover the deformation from point correspondences, as a 3D point cloud expressed in the coordinate frame of the camera which took the input image. In order to reconstruct the entire deformation to achieve applications such as augmented reality, one then reconstructs the function which maps points from the object's model to the deformed model. In most cases, the object's model has a 2D parameterization $\Omega \subset \mathbb{R}^2$, and the problem boils down to fitting a function $\varphi : \Omega \to \mathbb{R}^3$. This means that the target space has dimension three, and so $g = 3$. Therefore, the PRESS and Coupled-Measurement PRESS statistics are not equivalent and we only give the latter. The *spiderman* dataset was provided in (Bartoli et al., 2013) and comes with 1550 point correspondences. The *paper* dataset was provided in (Varol et al., 2012) and comes with 390 point correspondences. For both datasets, we reconstructed the 3D point cloud using the pointwise SfT method from (Bartoli et al., 2015). Each point being reconstructed independently, the point cloud is therefore quite noisy.

The *spiderman* dataset is an example of SfT data where the number of correspondences is rather high.

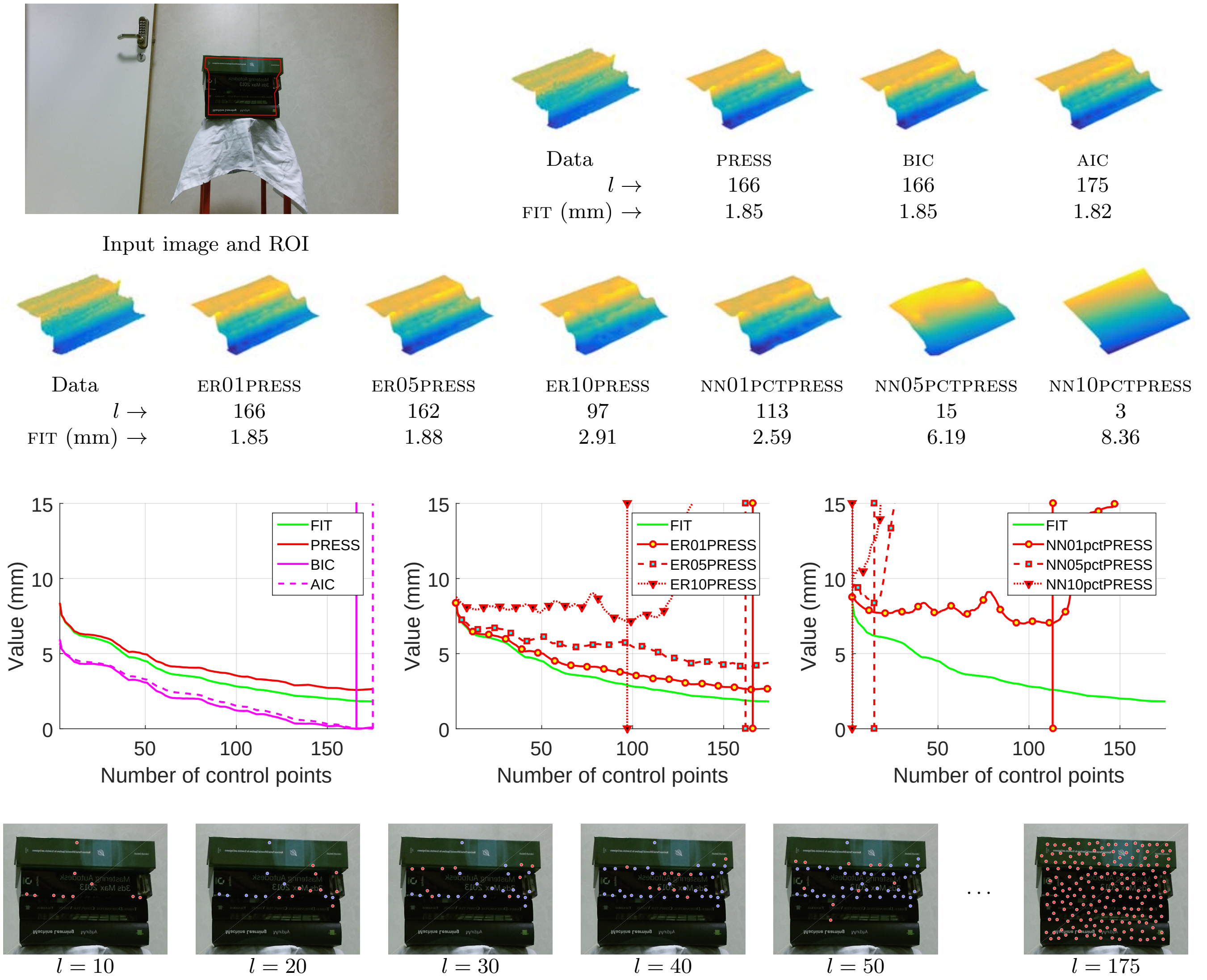


Figure 8: **Surface fitting for dense depth-maps: the sharp dataset.** The first two rows show the input image and the fitted surface for the different selection criteria. The third row shows the various selection criteria as a function of the number of control points. The y-axis of the graphs, labelled 'value', is in mm for all criteria as they give a value in depth units, except AIC and BIC. The fourth row shows the automatically placed control points in blue, with the last 10 placed points in red. The complete set is shown in red, right-most.

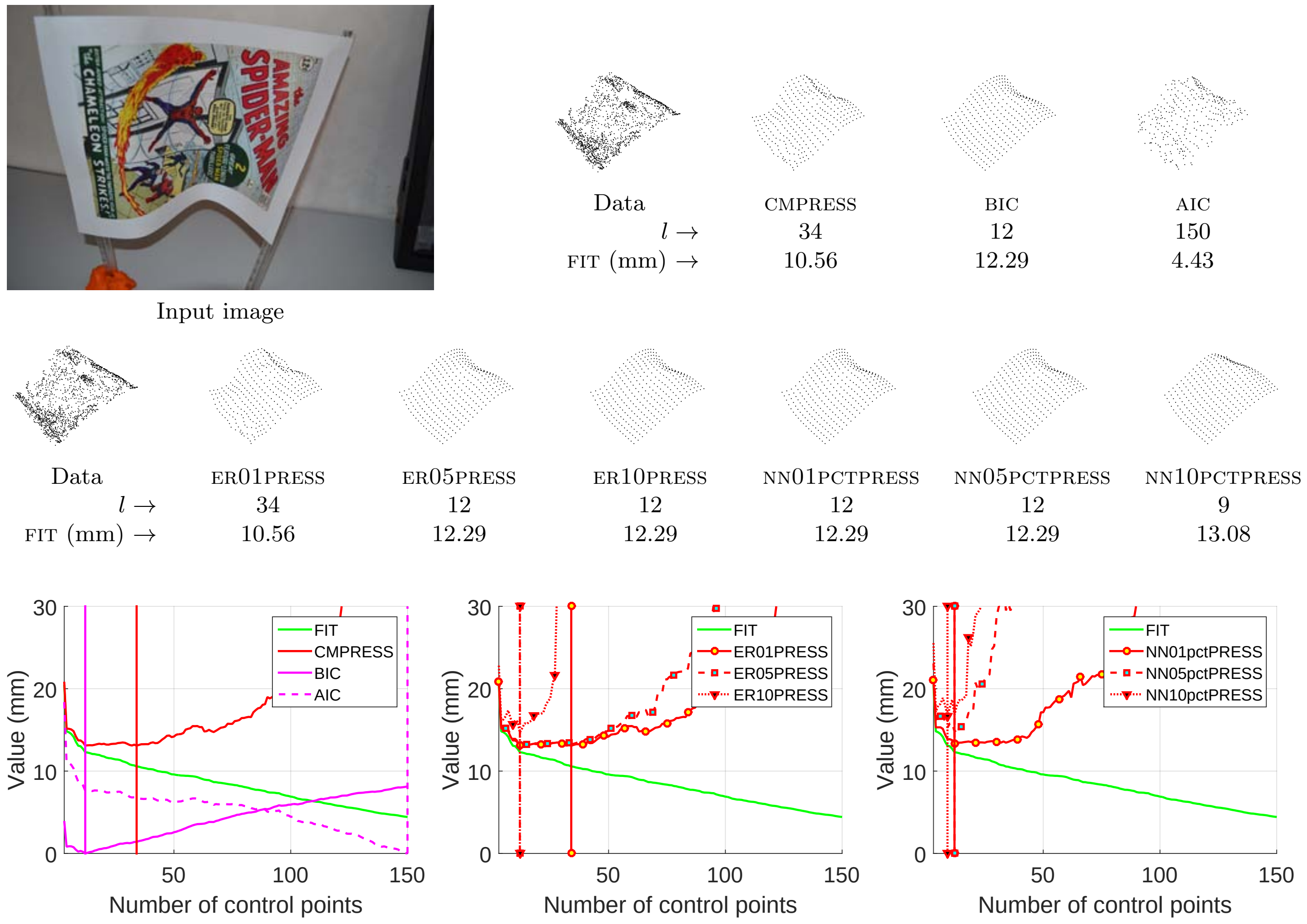


Figure 9: **Surface fitting for sparse depth-maps: the spiderman dataset.** The first two rows show the input image and the fitted surface for the different selection criteria. The third row shows the various selection criteria as a function of the number of control points. The y-axis of the graphs, labelled 'value', is in mm for all criteria as they give a value in 3D position units, except AIC and BIC.

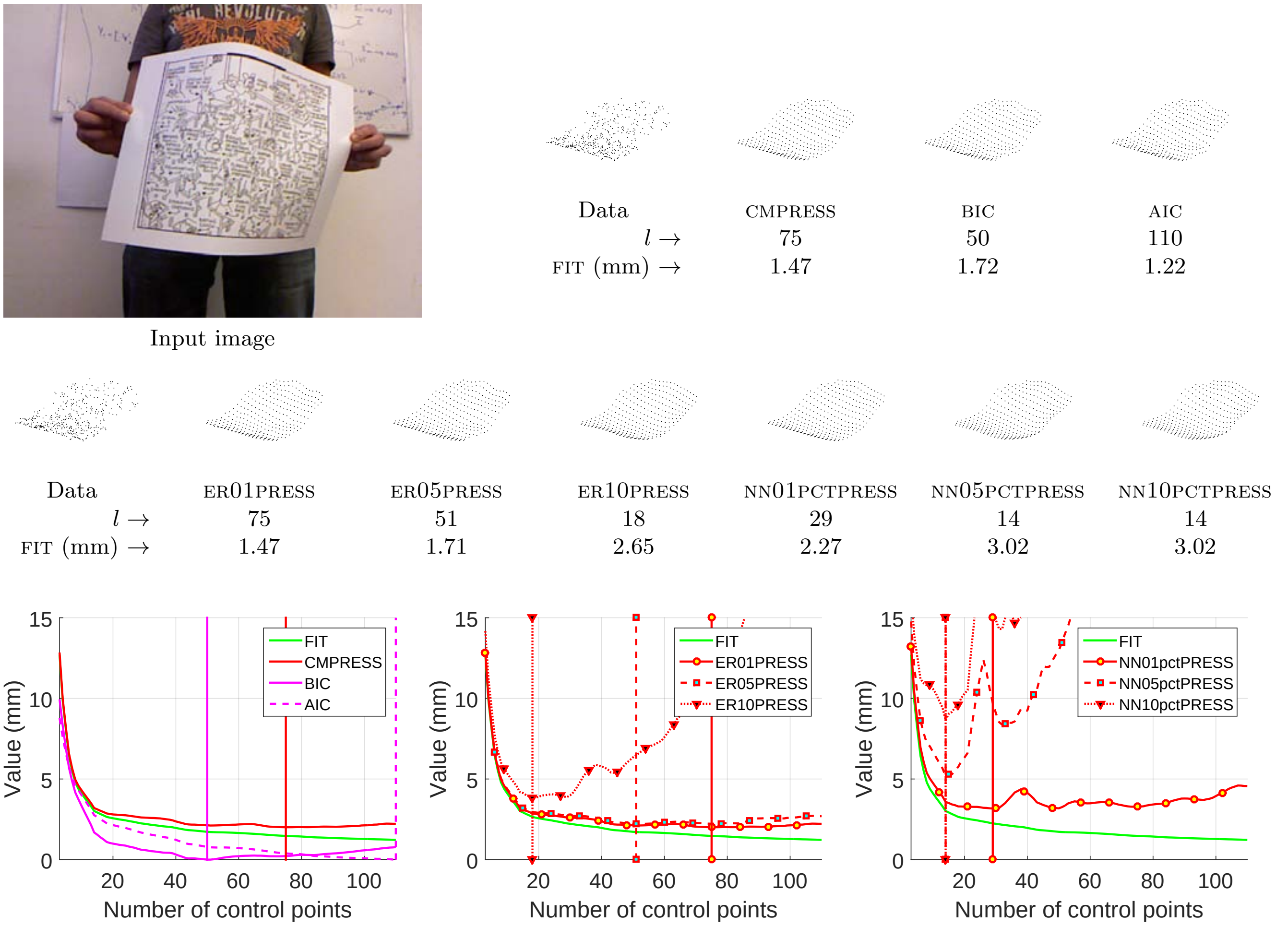
We set the maximum number of control points to 150. The input image and results are shown in figure 9. We observe significant differences between the reconstructed surfaces and that the fitting RMSR transitions from a steep to a milder decrease at 12 control points with a fitting RMSR of about 12 mm. Both lead to consistent observations in terms of the fitting quality. This transition is explained by the significant level of noise in the point cloud, in spite of the limited modeling error. We observe that NN10PCTPRESS underfits, as it selects only 9 control points. Eventhough it does not increase the fitting RMSR significantly, as it raises to 13.08 mm only, the reconstructed surface fails to capture the deformation's fine details. We observe that ER01PRESS and CMPRESS slightly overfit with 34 control points, a fitting RMSR reduced slightly to 10.56 mm, and a reconstructed surface affected by noise. Similarly, but to a much higher extent, we observe that AIC overfits, reducing the fitting RMSR significantly to 4.43 mm by selecting 150 control control, the maximum possible, and leading to an extremely noisy surface reconstruction. Finally, we observe that all the other criteria, namely BIC, ER05PRESS, ER10PRESS, NN01PCTPRESS and NN05PCTPRESS, select 12 control points, matching the reference number. They lead to the most visually convincing surface reconstruction.

The *paper* dataset is a typical example of SfT data, where the number of correspondences is low. We set the maximum number of control points to 110. The input image and results are shown in figure 10. We do not observe a clear transition point in the fitting RMSR, but rather a transition zone, located approximately between 10-20 control points. This is explained by the simple and smooth geometry of the observed surface. None of the criteria underfit. We observe that NN05PCTPRESS, NN10PCTPRESS and ER10PRESS all fall in the transition zone with 14, 14 and 18 control points selected, respectively. We then have NN01PCTPRESS, which also gave a visually satisfying surface reconstruction, with 29 control points selected. We observe that BIC and ER05PRESS overfit, with 50 and 51 control points selected and mild noise visible in the reconstructed surfaces, while ER01PRESS, CMPRESS and AIC clearly overfit, as visible from the reconstructed surfaces, with 75, 75 and 110 selected control points, respectively.

We draw the same overall observations as in the depth-map experiments case: most existing criteria lead to overfitting, while the proposed PRESS statistic based on an exclusion radius of 10% of the domain does not overfit.

6 Conclusion

We have proposed MEXPRESS, a formula which allows one to compute the PRESS statistic and many variants non-iteratively. PRESS uses the principle of Cross-Validation (CV) for Linear Least Squares regression. A usage of MEXPRESS is to compute PRESS, corresponding to Leave-One-Out CV, when multiple measurements are coupled, as in most regression problems. We have shown how to compute k-Fold PRESS and



Data	CMPRESS	BIC	AIC
l →	75	50	110
FIT (mm) →	1.47	1.72	1.22

Data	ER01PRESS	ER05PRESS	ER10PRESS	NN01PCTPRESS	NN05PCTPRESS	NN10PCTPRESS
l →	75	51	18	29	14	14
FIT (mm) →	1.47	1.71	2.65	2.27	3.02	3.02

Figure 10: **Surface fitting for sparse depth-maps: the paper dataset.** The first two rows show the input image and the fitted surface for the different selection criteria. The third row shows the various selection criteria as a function of the number of control points. The y-axis of the graphs, labelled 'value', is in mm for all criteria as they give a value in 3D position units, except AIC and BIC.

how to handle overfitting with Local-Exclusion PRESS. All these PRESS variants naturally handle coupled measurements. We have shown that MEXPRESS can be directly used to select the number of control points involved in linear fractional warps. These warps model perspective, and include the 2D homography as a special case. In this problem, the Local-Exclusion strategy turned out to be extremely important to cope with overfitting. We have also shown that MEXPRESS can be used to select the number of control points in surface fitting with dense and sparse data. Similarly, the Local-Exclusion strategy turned out to be efficient to mitigate overfitting. As a rule of thumb, we consistently found in our experiments that excluding data closest to the validation datum by 10% of the image's or domain's diagonal length, gives the most reliable results. However, we observed in some experiments that this strategy was slightly outperformed by excluding a fixed amount of nearest-neighbors of the validation datum. This conclusion is conditioned by the type of data being used, including the distribution of the data points and the variation in the transformation's curvature. In surface fitting, it is also conditioned by the strategy used to iteratively place the control points.

MEXPRESS opens the way to further research in model selection. A model's complexity can be represented by a varying number of parameters (such as the number of control points of an image warp) or the regularization weight used in the cost function. Being able to select the optimal number of parameters and regularization weight are unsolved problems. PRESS and Leave-One-Out CV are known to overfit but were computationally attractive for their non-iterative formulas. MEXPRESS brings flexibility, as it allows one to implement many different training and validation strategies non-iteratively. MEXPRESS may also allow one to address the question of robustness. Detecting blunders while fitting a flexible model is an open and tremendously difficult problem. This is because increasing the complexity of the model may make it fit the blunders to a good extent, leaving one with ambiguities regarding the data's 'blunderness'. Using the 'compatibility' between a datum and its neighbors may be used to collect evidence on the datum's blunderness. In this respect, MEXPRESS allows one to quickly and exhaustively assess the compatibility between a datum and various samples of its neighbors, and may form the basis for a robustified PRESS statistic, for which the notion of compatibility would be defined as the ability of neighboring data to predict one another.

Acknowledgements. I thank Mathias Gallardo and Ajad Chhatkuli for their help in creating the surface fitting datasets and the authors of (Varol et al., 2012) for the *paper* SfT dataset. This research has received funding from the EU's FP7 through the ERC research grant 307483 FLEXABLE.

A Proof of Proposition 1

The proof of proposition 1 follows the same steps as the proof for the PRESS (Bartoli, 2009). It requires the following lemma.

Lemma 1. *The model estimate $\bar{\mathbf{x}}_{(\mathcal{K})}$ obtained by holding out a set of measurements with index set $\mathcal{K} \in [1, m]$ is the same as the one obtained by replacing the $\mathcal{K}$ responses by their predictions with the model $\bar{\mathbf{x}}_{(\mathcal{K})}$:*

$$\bar{\mathbf{x}}_{(\mathcal{K})} \overset{\text{def}}{=} \left(\mathtt{I}_{(\mathcal{K})}\mathtt{A}\right)^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b} = \mathtt{A}^{\dagger}\tilde{\mathbf{b}}_{(\mathcal{K})} \quad \text{with} \quad \tilde{\mathbf{b}}_{(\mathcal{K})} \overset{\text{def}}{=} \mathtt{I}_{(\mathcal{K})}\mathbf{b} + \mathtt{I}_{(-\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})}, \tag{22}$$

where we used the notation $\mathtt{I}_{(\mathcal{K})}$ for an identity matrix whose diagonal entries with index in $\mathcal{K}$ are put to zero, and $\mathtt{I}_{(-\mathcal{K})} \overset{\text{def}}{=} \mathtt{I} - \mathtt{I}_{(\mathcal{K})}$. Formally, $\mathtt{I}_{(\mathcal{K})} = \operatorname{diag}(\mathbf{d})$ with $\mathbf{d}_{\mathcal{K}} = 0$ and $\mathbf{d}_{-\mathcal{K}} = 1$.

Proof of lemma 1. Using the definition of $\tilde{\mathbf{b}}_{(\mathcal{K})}$ we have:

$$\mathtt{A}^{\dagger}\tilde{\mathbf{b}}_{(\mathcal{K})} = \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b} + \mathtt{A}^{\dagger}\mathtt{I}_{(-\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})} \tag{23}$$

Because $\mathtt{I}_{(-\mathcal{K})} = \mathtt{I} - \mathtt{I}_{(\mathcal{K})}$ this may be expanded as:

$$\mathtt{A}^{\dagger}\tilde{\mathbf{b}}_{(\mathcal{K})} = \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b} + \mathtt{A}^{\dagger}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})} - \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})}. \tag{24}$$

The second term is simplified to $\mathtt{A}^{\dagger}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})} = \bar{\mathbf{x}}_{(\mathcal{K})}$. Using the definition of $\bar{\mathbf{x}}_{(\mathcal{K})}$, the third term is simplified to $\mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})} = \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b}$. The overall expression is thus simplified to:

$$\mathtt{A}^{\dagger}\tilde{\mathbf{b}}_{(\mathcal{K})} = \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b} + \bar{\mathbf{x}}_{(\mathcal{K})} - \mathtt{A}^{\dagger}\mathtt{I}_{(\mathcal{K})}\mathbf{b} = \bar{\mathbf{x}}_{(\mathcal{K})}, \tag{25}$$

concluding the proof. □

Proof of proposition 1. The case of an empty set $\mathcal{K}$ is straightforward as then $\mathbf{e}_{(\mathcal{K})}$ is a zero-length vector. We start by rewriting the predictions in $\mathcal{K}$ by the model fitted with all measurements as $\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}} = \mathtt{A}_{\mathcal{K}}\mathtt{A}^{\dagger}\mathbf{b} = \hat{\mathtt{A}}_{\mathcal{K}}\mathbf{b}$. Similarly, we rewrite the predictions in $\mathcal{K}$ by the model fitted with all but the measurements in $\mathcal{K}$ as $\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} = \mathtt{A}_{\mathcal{K}}\mathtt{A}^{\dagger}\tilde{\mathbf{b}}_{(\mathcal{K})} = \hat{\mathtt{A}}_{\mathcal{K}}\tilde{\mathbf{b}}_{(\mathcal{K})}$, where the first equality is given by lemma 1. The difference between the two predictions is thus given by:

$$\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}} - \mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} = \hat{\mathtt{A}}_{\mathcal{K}}\left(\mathbf{b} - \tilde{\mathbf{b}}_{(\mathcal{K})}\right) = \hat{\mathtt{A}}_{\mathcal{K}}\left(\mathbf{b} - \mathtt{I}_{(\mathcal{K})}\mathbf{b} - \mathtt{I}_{(-\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})}\right), \tag{26}$$

where the second equality is obtained by using the definition of $\tilde{\mathbf{b}}_{(\mathcal{K})}$. We then use $\mathtt{I} - \mathtt{I}_{(\mathcal{K})} = \mathtt{I}_{(-\mathcal{K})}$, leading to:

$$\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}} - \mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} \;=\; \hat{\mathtt{A}}_{\mathcal{K}}\left(\mathtt{I}_{(-\mathcal{K})}\mathbf{b} - \mathtt{I}_{(-\mathcal{K})}\mathtt{A}\bar{\mathbf{x}}_{(\mathcal{K})}\right). \tag{27}$$

By noting that $\hat{\mathtt{A}}_{\mathcal{K}}\mathtt{I}_{(-\mathcal{K})}\mathtt{A} = \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\mathtt{A}_{\mathcal{K}}$ and $\hat{\mathtt{A}}_{\mathcal{K}}\mathtt{I}_{(-\mathcal{K})}\mathbf{b} = \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\mathbf{b}_{\mathcal{K}}$ we rearrange the equation as:

$$(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}})\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} + \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\mathbf{b}_{\mathcal{K}} \;=\; \mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}. \tag{28}$$

We subtract $\mathbf{b}_{\mathcal{K}}$ to each side of the equation:

$$\left(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\right)\left(\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} - \mathbf{b}_{\mathcal{K}}\right) \;=\; \mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}} - \mathbf{b}_{\mathcal{K}}, \tag{29}$$

and arrive at:

$$\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}}_{(\mathcal{K})} - \mathbf{b}_{\mathcal{K}} \;=\; \left(\mathtt{I} - \hat{\mathtt{A}}_{\mathcal{K},\mathcal{K}}\right)^{-1}\left(\mathtt{A}_{\mathcal{K}}\bar{\mathbf{x}} - \mathbf{b}_{\mathcal{K}}\right), \tag{30}$$

concluding the proof. □

References

H. Akaike. A new look at the statistical model identification. IEEE *Transactions on Automation and Control*, AC-19(6):716–723, December 1974.

D. M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3): 469–475, August 1971.

A. Bartoli. Maximizing the predictivity of smooth deformable image warps through cross-validation. *Journal of Mathematical Imaging and Vision*, 31(2-3):133–145, July 2008.

A. Bartoli. On computing the prediction sum of squares statistic in linear least squares problems with multiple parameter or measurement sets. *International Journal of Computer Vision*, 85(2):133–142, November 2009.

A. Bartoli, M. Perriollat, and S. Chambon. Generalized thin-plate spline warps. *International Journal of Computer Vision*, 88(1):85–110, May 2010.

A. Bartoli, D. Pizarro, and T. Collins. A robust analytical solution to isometric shape-from-template with focal length calibration. In *International Conference on Computer Vision*, 2013.

A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-template. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2099–2118, October 2015.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

F. L. Bookstein. Principal warps: Thin-Plate Splines and the decomposition of deformations. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989.

L. Breiman. Fitting additive models to regression data. *Computational Statistics and Data Analysis*, 15: 13–46, 1993.

F. Brunet, A. Bartoli, R. Malgouyres, and N. Navab. NURBS warps. In *British Machine Vision Conference*, 2009.

P. Dierckx. An algorithm for surface-fitting with spline functions. *IMA Journal of Numerical Analysis*, 1 (3):267–283, 1981.

P. Dierckx. *Curve and surface fitting with splines.* Oxford University Press, 1993.

D. Forsyth and J. Ponce. *Computer Vision – A Modern Approach.* Prentice Hall, second edition edition, 2003.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2003. Second Edition.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, 2001.

D. L. B. Jupp. Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, April 1978.

F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *International Conference on Computer Vision*, 2005.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

S. Miyata and X. Shen. Free-knot splines and adaptive knot selection. *Journal of the Japanese Statistical Society*, 35(2):303–324, 2005.

N. Molinari, J.-F. Durand, and R. Sabatier. Bounded optimal knots for regression splines. *Computational Statistics and Data Analysis*, 45:159–178, 2004.

N. Nguyen, P. Milanfar, and G. Golub. Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 10(9):1299–1308, September 2001.

M. Perriollat and A. Bartoli. A computational model of bounded developable surfaces with application to image-based 3D reconstruction. *Computer Animation and Virtual Worlds*, 24(5):459–476, September 2013.

M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *International Journal of Computer Vision*, 95(2):124–137, November 2011.

D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registation using Free-Form Deformations: Application to breast MR images. IEEE *Transactions on Medical Imaging*, 18(8):712–721, August 1999.

M. Salzmann, J. Pilet, S. Ilic, and P. Fua. Surface deformation models for nonrigid 3D shape recovery. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1–7, August 2007.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001.

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

L. Sevilla-Lara and E. Learned-Miller. Distribution fields for tracking. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

F. Steinbrücker, C. Kerl, J. Sturm, and D. Cremers. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In *International Conference on Computer Vision*, 2013.

J. Süssmuth, Q. Meyer, and G. Greiner. Surface reconstruction based on hierarchical floating radial basis functions. *Computer Graphics Forum*, 29(6):1854–1864, 2010.

H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

W. Xu, M. Salzmann, Y. Wang, and Y. Liu. Nonrigid surface registration and completion from RGBD images. In *European Conference on Computer Vision*, 2014.

X. Yan and X. G. Su. *Linear Regression Analysis: Theory and Computing.* World Scientific Publishing, 2009. ISBN 9812834109.

M. Zollhöfer, M. Niessner, S. Izadi, C. Rhemann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics*, 33(4), 2014.