On Computing the Prediction Sum of Squares Statistic in Linear Least Squares Problems with Multiple Parameter or Measurement Sets

Adrien Bartoli

LASMEA (CNRS / Université Blaise Pascal), Clermont-Ferrand, France

Adrien.Bartoli@gmail.com

www.lasmea.univ-bpclermont.fr/Personnel/Adrien.Bartoli

Abstract

The prediction sum of squares is a useful statistic for comparing different models. It is based on the principle of leave-one-out or ordinary cross-validation, whereby every measurement is considered in turn as a test set, for the model parameters trained on all but the held out measurement. As for linear least squares problems, there is a simple well-known non-iterative formula to compute the prediction sum of squares without having to refit the model as many times as the number of measurements. We extend this formula to cases where the problem has multiple parameter or measurement sets.

We report experimental results on the fitting of a warp between two images, for which the number of deformation centres is automatically selected, based on one of the proposed non-iterative formulae.

Keywords: PRESS, cross-validation, registration, image warp, deformation centre, Thin-Plate Spline.

1 Introduction

The prediction sum of squares (PRESS) is a statistic based on the leave-one-out technique. It was proposed by Allen in 1974 [1], and is typically used to compare different models. It is equivalent to the sum of studentized residuals, and can be extended to select parameters such as the regularization weight in smoothing splines, as shown by Wahba *et al.* [12]. The PRESS is a statistic that depends on a chosen cost function, and is in a sense complementary to this cost function. The cost function often expresses the discrepancy between measurements and the values predicted by a parametric model. While minimizing the cost function allows one to find the model parameters, it is clear that the most complex model always has the lowest residual error. In other words, the 'best' model cannot be selected based on the residual error only. The PRESS statistic, however, does not depend on some particular model parameters, but on the model itself. As with techniques based on cross-validation, it expresses to which extent a particular model is able to generalize to new data. The PRESS should therefore be used as a measure of *predictivity* to compare and select the 'best' model, while minimizing the cost function gives the parameters of a particular model.

One of the main problems of cross-validation techniques is their computational cost. For the case of regular linear least squares, it is well-known that there is a simple non-iterative formula giving the PRESS without having to solve as many problems as there are measurements[1] [6]. This has been derived for several variants of the basic linear least squares problem. For instance, Tarpey [11] examines the case of restricted least squares.

We derive non-iterative PRESS formulae for those cases with multiple parameter or measurement sets. Multiple parameter sets means that the same measurement has several predictions, while multiple measurement sets means that a prediction is compared to several measurements. We assume that the model is well constrained by the measurements, or in other words, that the design matrix in the linear least squares system of equations has full rank. The case of regularized least squares is left as future work. We report experimental results showing how one of the proposed non-iterative PRESS formulae can be used to assess the fit of an inter-image warp we proposed in [3], and to automatically select the number of deformation centres for this warp. In this case, the warp with different numbers of deformation centres gives the different models to be compared. The 'best' model, and thus the 'best' number of deformation centres, is selected as the one which minimizes the PRESS statistic.

Notation. In general, we write vectors in bold fonts, *e.g.* $\mathbf{x}$, matrices in sans-serif and calligraphic fonts, *e.g.* A, $\mathcal{W}$, and scalars in italics, *e.g.* m. Matrix inverse is written as in A^{-1}. Matrix and vector transpose are written as in $\mathsf{A}^\top$ and $\mathbf{x}^\top$. The homogeneous coordinates of a 2D point $\mathbf{q}$ are written as a (3×1) vector overlaid with a tilde, as in $\tilde{\mathbf{q}}$.

Organization of the article. We review the non-iterative PRESS formula for standard linear least squares in §2. The extension to linear least squares problems with multiple parameter or measurement sets is given in §3. The proofs of some key results are reported in §4. An application to estimating an image warp is eventually given in §5, and our conclusion in §6.

2 Background: Standard Linear Least Squares

Let $\mathbf{x}$ be the *parameter vector*, A the design matrix with m rows $\mathbf{a}_j$, $j = 1, \ldots, m$ and $\mathbf{b}$ the $(m \times 1)$ *measurement vector*. The j-th measurement b_j is thus predicted by the model as $\mathbf{a}_j^\top \mathbf{x}$. Consider a regular linear least squares cost function:

$$\mathcal{E}^2_{\text{STD}}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{j=1}^{m} \left(\mathbf{a}_j^\top \mathbf{x} - b_j \right)^2 = \frac{1}{m} \|\mathsf{A}\mathbf{x} - \mathbf{b}\|_2^2,$$

where $\|\mathbf{u}\|_2$ is the vector two-norm *i.e.*, $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}^\top \mathbf{u}}$. The solution to the problem $\min_{\mathbf{x}} \mathcal{E}^2_{\text{STD}}(\mathbf{x})$ is:

$$\bar{\mathbf{x}} = \mathsf{A}^\dagger \mathbf{b},$$

[1] 'Non-iterative formula' emphasizes the fact that explicitly training the model with all but one measurement and testing on this measurement is not required.

with $\mathsf{A}^\dagger$ the matrix pseudo-inverse. The PRESS $\mathcal{P}_{\text{STD}}$ is defined as follows. The model is fitted without the j-th measurement giving the parameter vector $\bar{\mathbf{x}}_{(j)}$ and is used to predict the j-th measurement as $\mathbf{a}_j^\top \bar{\mathbf{x}}_{(j)}$. This prediction is compared against the actual measurement b_j. This is averaged over the m measurements, giving:

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

Directly using this formula for estimating the PRESS would be extremely inefficient since the model has to be fitted m times to compute all the $\bar{\mathbf{x}}_{(j)}$. However, it is well-known that there is a non-iterative formula giving the PRESS as:

$$\mathcal{P}^2_{\text{STD}} = \frac{1}{m}\left\|\Gamma\left(\mathrm{I}-\hat{\mathsf{A}}\right)^{-1}\left(\hat{\mathsf{A}}-\mathrm{I}\right)\mathbf{b}\right\|_2^2, \tag{1}$$

with $\hat{\mathsf{A}} = \mathsf{A}\mathsf{A}^\dagger$ the *hat matrix*, I the identity matrix, and where $\Gamma(\mathsf{B})$ keeps only the diagonal entries of the square matrix B. Note that $\left(\hat{\mathsf{A}}-\mathrm{I}\right)\mathbf{b} = \mathsf{A}\bar{\mathbf{x}} - \mathbf{b}$ is the residual vector. A proof of formula (1) is given in *e.g.* [2, 10].

3 Multiple Parameter or Measurement Sets

This section brings our main results. Proofs are reported in §4. We deal with four cases of parameter and measurement sets. Standard linear least squares is the first case, and corresponds to a single parameter set and a single measurement set. In this case, the model parameters and the measurements are both contained in a vector, respectively $\mathbf{x}$ and $\mathbf{b}$. In the multiple parameter sets case, the model parameters are contained in a matrix L, each column of which being a parameter set. The single measurement set $\mathbf{b}$ is replicated using an outer vector product as $\mathbf{b}\boldsymbol{\omega}^\top$, where $\boldsymbol{\omega}$ is an $(n \times 1)$ vector that scales the replicated measurements. In the multiple measurement sets case, each prediction matches multiple measurements in a matrix R. The single parameter set is replicated using an outer vector product as $\mathbf{x}\boldsymbol{\omega}^\top$.

Our basic assumption to formulate the PRESS for those different cases, is that the measurements lying on the same row of the 'measurement matrix' R (whether replicated from a single measurement set or not) are *linked*, and should be held out simultaneously.

Multiple parameter sets, single and multiple measurement sets. We first deal with the multiple parameter and measurement sets case (MPM). This kind of linear least squares problems has n sets of parameters and measurements represented in matrix form: L is the parameter matrix and R is the measurement matrix with rows $\mathbf{r}_j$. They both have n columns, each being respectively a parameter and a measurement set, the former linked to the latter through the design matrix A. The least squares cost function is as follows:

$$\mathcal{E}^2_{\text{MPM}}(\mathsf{L}) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{a}_j^\top \mathsf{L} - \mathbf{r}_j^\top\right\|_2^2 = \frac{1}{m}\|\mathsf{A}\mathsf{L}-\mathsf{R}\|_{\mathcal{F}}^2,$$

where $\|\mathsf{U}\|_{\mathcal{F}}$ is the matrix Frobenius norm *i.e.*, $\|\mathsf{U}\|_{\mathcal{F}} \stackrel{\text{def}}{=} \sqrt{\text{tr}(\mathsf{U}^\top \mathsf{U})}$. The solution to the problem $\min_{\mathsf{L}} \mathcal{E}^2_{\text{MPM}}(\mathsf{L})$ is:

$$\bar{\mathsf{L}} \stackrel{\text{def}}{=} \mathsf{A}^\dagger \mathsf{R}.$$

The PRESS is defined by:

$$\mathcal{P}^2_{\text{MPM}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{j=1}^{m} \left\| \mathbf{a}_j^\top \bar{\mathsf{L}}_{(j)} - \mathbf{r}_j^\top \right\|_2^2,$$

and can be computed efficiently with the following non-iterative formula:

$$\mathcal{P}^2_{\text{MPM}} = \frac{1}{m} \left\| \Gamma \left(\mathrm{I} - \hat{\mathsf{A}} \right)^{-1} \left(\hat{\mathsf{A}} - \mathrm{I} \right) \mathsf{R} \right\|_{\mathcal{F}}^2, \tag{2}$$

which is exactly as (1) for the standard linear least squares case, except that the vector two-norm is replaced by the matrix Frobenius norm. This is demonstrated very easily by following the proof in [2, 10], replacing the vector by the matrix norm. The intuition is that each column of R is independent, in the sense that the corresponding parameters lie in a different column in L, and that $\|\mathsf{U}\|^2_{\mathcal{F}} = \|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 + \ldots$, where $\mathbf{u}_1$, $\mathbf{u}_2$, $\ldots$, are the columns[2] of matrix U. The problem can thus be split into n standard linear least squares problems, and their PRESS combined together to give (2).

The case of multiple parameter sets and single measurement set in $\mathbf{b}$ is a special case of the above derivation. Formula (2) holds, replacing the measurement matrix R by the replicated measurement vector $\mathbf{b}\boldsymbol{\omega}^\top$, where $\boldsymbol{\omega}$ contains scaling factors.

In the above definition of the PRESS, we consider that the rows of the measurement matrix must be held out simultaneously. For the multiple parameter sets case, holding out each measurement independently makes little difference, since the independence properties of the parameters are preserved. Only the normalization factor would change in the PRESS formula, from $\frac{1}{m}$ to $\frac{1}{nm}$.

Single parameter set, multiple measurement sets. We investigate the case where there is a single parameter vector with multiple measurement sets (MM). In other words, each model prediction matches several measurements. This is modeled by the following cost function:

$$\mathcal{E}^2_{\text{MM}}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{m} \|\mathsf{C}\mathsf{L} - \mathsf{R}\|^2_{\mathcal{F}} \qquad \text{with} \qquad \mathsf{L} = \mathbf{x}\mathbf{v}^\top,$$

where C is the m row design matrix, and $\mathbf{v}$ is the 'all-one' vector *i.e.*, $\mathbf{v}^\top = (1 \ \cdots \ 1)$. This is a particular case of the multiple scaled measurement sets described below. We can obviously not apply the standard PRESS formula[3] since n linked measurements must be removed jointly. Holding out only one measurement at a time underestimates

[2] This obviously also holds with the rows.

[3] This formula is (1) and could be direcly (but incorrectly) applied to the 'vectorized' formulation (8) of the problem.

the PRESS. The difference with the true PRESS is more involved than just a scale factor as in the above multiple parameter sets case, since the parameters are dependent, and their estimated values would change.

In the next paragraph, we deal with a more general case, whereby each measurement set is scaled differently. By setting the $(n \times 1)$ scaling vector $\boldsymbol{\omega}$ to the all-one vector $\mathbf{v}$, and noting that $\|\mathbf{v}\|_2^2 = n$ the solution to $\min_{\mathbf{x}} \mathcal{E}^2_{\mathrm{MM}}(\mathbf{x})$ is obtained from the general solution (5) as:

$$\bar{\mathbf{x}} \;=\; \frac{1}{n}\mathsf{C}^{\dagger}\mathsf{R}\mathbf{v}.$$

The PRESS is defined by:

$$\mathcal{P}^2_{\mathrm{MM}} \;\stackrel{\mathrm{def}}{=}\; \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{c}_j^{\top}\bar{\mathbf{x}}_{(j)}\mathbf{v}^{\top} - \mathbf{r}_j^{\top}\right\|^2,$$

with $\mathbf{c}_j$ and $\mathbf{r}_j$ the rows of C and R respectively. The non-iterative PRESS formula we derive is:

$$\mathcal{P}^2_{\mathrm{MM}} \;=\; \frac{1}{m}\left\|\Gamma\left(\mathrm{I}-\hat{\mathsf{C}}\right)^{-1}\left(\frac{1}{n}\hat{\mathsf{C}}\mathsf{R}\mathsf{V} - \mathsf{R} + \Gamma\left(\hat{\mathsf{C}}\right)\mathsf{R}\left(\mathrm{I}-\frac{1}{n}\mathsf{V}\right)\right)\right\|^2_{\mathcal{F}}, \tag{3}$$

with $\mathsf{V} = \mathbf{v}\mathbf{v}^{\top}$ the 'all-one' $(n \times n)$ matrix. Specializing the general, scaled measurement equation (7) with $\boldsymbol{\omega} = \mathbf{v}$ to get (3) is straightforward.

Single parameter set, multiple scaled measurement sets. This case generalizes the previous one by incorporating a different scale for each of the measurement sets (MSM) *i.e.*, for each column in R, through an $(n \times 1)$ scaling vector $\boldsymbol{\omega}$:

$$\mathcal{E}^2_{\mathrm{MSM}}(\mathbf{x}) \;\stackrel{\mathrm{def}}{=}\; \frac{1}{m}\|\mathsf{C}\mathsf{L} - \mathsf{R}\|^2_{\mathcal{F}} \qquad \text{with} \qquad \mathsf{L} \;=\; \mathbf{x}\boldsymbol{\omega}^{\top}. \tag{4}$$

The solution to $\min_{\mathbf{x}} \mathcal{E}^2_{\mathrm{MSM}}(\mathbf{x})$ is:

$$\bar{\mathbf{x}} \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^{\dagger}\mathsf{R}\boldsymbol{\omega}. \tag{5}$$

The PRESS is defined by:

$$\mathcal{P}^2_{\mathrm{MSM}} \;\stackrel{\mathrm{def}}{=}\; \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{c}_j^{\top}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^{\top} - \mathbf{r}_j^{\top}\right\|_2^2, \tag{6}$$

and we demonstrate below that the non-iterative PRESS formula is:

$$\mathcal{P}^2_{\mathrm{MSM}} \;=\; \frac{1}{m}\left\|\Gamma\left(\mathrm{I}-\hat{\mathsf{C}}\right)^{-1}\left(\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathsf{C}}\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^{\top} - \mathsf{R} + \Gamma\left(\hat{\mathsf{C}}\right)\mathsf{R}\left(\mathrm{I}-\frac{1}{\|\boldsymbol{\omega}\|_2^2}\boldsymbol{\omega}\boldsymbol{\omega}^{\top}\right)\right)\right\|^2_{\mathcal{F}}. \tag{7}$$

This looks like the usual direct solution (2) except that an extra 'corrective' term $\Gamma\left(\hat{\mathsf{C}}\right)\mathsf{R}\left(\mathrm{I}-\frac{1}{\|\boldsymbol{\omega}\|_2^2}\boldsymbol{\omega}\boldsymbol{\omega}^{\top}\right)$ is added to the residual matrix $\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathsf{C}}\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^{\top} - \mathsf{R}$.

4 Proofs for the Multiple Scaled Measurement Sets Case

The solution to $\min_{\mathbf{x}} \mathcal{E}^2_{\mathbf{MSM}}(\mathbf{x})$. We start by deriving the solution $\bar{\mathbf{x}}$ in equation (5). This equation means that $\bar{\mathbf{x}}$ is the $\boldsymbol{\omega}$ weighted average of the solutions for each set of measurements. It is derived by rewriting the cost function (4) as:

$$\mathcal{E}^2_{\mathrm{MSM}}(\mathbf{x}) \;=\; \frac{1}{m}\|\boldsymbol{\omega} \otimes \mathsf{C}\mathbf{x} - \mathbf{r}\|_2^2, \tag{8}$$

where $\mathbf{r} = \mathrm{vect}(\mathsf{R})$ is the column-wise vectorization of R, and $\otimes$ is the Kronecker product. From this rewriting, we obtain:

$$\bar{\mathbf{x}} \;=\; (\boldsymbol{\omega} \otimes \mathsf{C})^{\dagger}\mathbf{r},$$

which rewrites as:

$$\bar{\mathbf{x}} \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\left(\boldsymbol{\omega}^{\top} \otimes \mathsf{C}^{\dagger}\right)\mathbf{r} \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^{\dagger}\mathsf{R}\boldsymbol{\omega}.$$

The non-iterative PRESS formula $\mathcal{P}^2_{\mathbf{MSM}}$. The next step is to derive the non-iterative PRESS formula (7). The proof follows similar steps as the proof for the basic non-iterative PRESS formula. We start from the PRESS definition (6). Defining $\mathbf{e}_j$ as a zero vector with one at the j-th element and $\mathsf{D}_j \stackrel{\mathrm{def}}{=} \mathrm{I} - \Delta(\mathbf{e}_j)$, where Δ constructs a diagonal matrix from a vector, we have:

$$\bar{\mathbf{x}}_{(j)} \;\stackrel{\mathrm{def}}{=}\; \arg\min_{\mathbf{x}} \|\mathsf{D}_j\mathsf{C}\mathbf{x}\boldsymbol{\omega}^{\top} - \mathsf{D}_j\mathsf{R}\|_{\mathcal{F}}^2.$$

Note that matrix D_j has the properties:

$$\begin{aligned} \mathsf{D}_j\mathsf{D}_j &= \mathsf{D}_j \\ \mathsf{D}_j^{\top} &= \mathsf{D}_j \\ \mathrm{I} - \mathsf{D}_j &= \Delta(\mathbf{e}_j). \end{aligned}$$

Similarly as for the solution (5), we get:

$$\bar{\mathbf{x}}_{(j)} \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}(\mathsf{D}_j\mathsf{C})^{\dagger}\mathsf{D}_j\mathsf{R}\boldsymbol{\omega}. \tag{9}$$

The lemma we demonstrate in the next paragraph states that:

$$\bar{\mathbf{x}}_{(j)} \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^{\dagger}\tilde{\mathsf{R}}_j\boldsymbol{\omega}, \tag{10}$$

with $\tilde{\mathsf{R}}_j$ the measurement matrix R with the j-th row replaced by its prediction $\mathbf{c}_j^{\top}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^{\top}$:

$$\tilde{\mathsf{R}}_j \;\stackrel{\mathrm{def}}{=}\; \mathsf{D}_j\mathsf{R} + (\mathrm{I} - \mathsf{D}_j)\mathsf{C}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^{\top}. \tag{11}$$

We note that:

$$(\mathrm{I} - \mathsf{D}_j)\mathsf{C}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top \;=\; \Delta(\mathbf{e}_j)\mathsf{C}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top \;=\; \mathbf{e}_j\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top.$$

The prediction of the j-th data with the global model $\bar{\mathbf{x}}$ is $\mathbf{c}_j^\top\bar{\mathbf{x}}\boldsymbol{\omega}^\top$. By substituting $\bar{\mathbf{x}}$ from equation (5), we get:

$$\mathbf{c}_j^\top\bar{\mathbf{x}}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathbf{c}_j^\top\mathsf{C}^\dagger\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^\top, \tag{12}$$

where $\hat{\mathbf{c}}_j$ is the j-th row of the hat matrix $\hat{\mathsf{C}} = \mathsf{C}\mathsf{C}^\dagger$. Similarly, we rewrite the prediction of the j-th data with the partial model $\bar{\mathbf{x}}_{(j)}$ from equation (10) as:

$$\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\tilde{\mathsf{R}}_{(j)}\boldsymbol{\omega}\boldsymbol{\omega}^\top. \tag{13}$$

Taking the difference between the two predictions as rewritten in equations (12) and (13), and factorizing gives:

$$\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^\top - \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\tilde{\mathsf{R}}_j\boldsymbol{\omega}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\left(\mathsf{R} - \tilde{\mathsf{R}}_j\right)\boldsymbol{\omega}\boldsymbol{\omega}^\top.$$

Using (11), we substitute $\tilde{\mathsf{R}}_j = \mathsf{D}_j\mathsf{R} + \mathbf{e}_j\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top$ which gives:

$$\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\left(\mathsf{R} - \mathsf{D}_j\mathsf{R} - \mathbf{e}_j\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top\right)\boldsymbol{\omega}\boldsymbol{\omega}^\top,$$

which, simplifying $\mathsf{R} - \mathsf{D}_j\mathsf{R} = \mathbf{e}_j\mathbf{r}_j^\top$ and since $\boldsymbol{\omega}^\top\boldsymbol{\omega} = \|\boldsymbol{\omega}\|_2^2$, transforms to:

$$\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathbf{c}}_j^\top\left(\mathbf{e}_j\mathbf{r}_j^\top\boldsymbol{\omega} - \|\boldsymbol{\omega}\|_2^2\mathbf{e}_j\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\right)\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{c}_{j,j}\left(\mathbf{r}_j^\top\boldsymbol{\omega} - \|\boldsymbol{\omega}\|_2^2\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\right)\boldsymbol{\omega}^\top,$$

where $\hat{c}_{j,j}$ is the j-th diagonal element of the hat matrix $\hat{\mathsf{C}}$. We thus have rewritten the original prediction difference as follows:

$$\mathbf{c}_j^\top\bar{\mathbf{x}}\boldsymbol{\omega}^\top - \mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{c}_{j,j}\left(\mathbf{r}_j^\top\boldsymbol{\omega} - \|\boldsymbol{\omega}\|_2^2\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\right)\boldsymbol{\omega}^\top.$$

Rearranging the terms gives:

$$\mathbf{c}_j^\top\bar{\mathbf{x}}\boldsymbol{\omega}^\top \;=\; \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{c}_{j,j}\mathbf{r}_j^\top\boldsymbol{\omega}\boldsymbol{\omega}^\top + (1 - \hat{c}_{j,j})\,\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top.$$

Adding $\mathbf{r}_j^\top\left(\hat{c}_{j,j}\mathrm{I} - \mathrm{I} - \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{c}_{j,j}\boldsymbol{\omega}\boldsymbol{\omega}^\top\right)$ on both sides gives:

$$\mathbf{c}_j^\top\bar{\mathbf{x}}\boldsymbol{\omega}^\top + \mathbf{r}_j^\top\left(\hat{c}_{j,j}\mathrm{I} - \mathrm{I} - \frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{c}_{j,j}\boldsymbol{\omega}\boldsymbol{\omega}^\top\right) \;=\; (1 - \hat{c}_{j,j})\left(\mathbf{c}_j^\top\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top - \mathbf{r}_j^\top\right),$$

from which:

$$\mathbf{c}_j^\top \bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top - \mathbf{r}_j^\top \;=\; \frac{1}{1-\hat{c}_{j,j}}\left(\mathbf{c}_j^\top \bar{\mathbf{x}}\boldsymbol{\omega}^\top - \mathbf{r}_j^\top + \mathbf{r}_j^\top \hat{c}_{j,j}\left(\mathrm{I} - \frac{1}{\|\boldsymbol{\omega}\|_2^2}\boldsymbol{\omega}\boldsymbol{\omega}^\top\right)\right).$$

We observe that $\mathbf{c}_j^\top \bar{\mathbf{x}}\boldsymbol{\omega}^\top - \mathbf{r}_j^\top$ is the residual vector for the j-th measurement. Replacing $\bar{\mathbf{x}}$ from equation (5) and summing the squared norm over j, we get the non-iterative PRESS formula:

$$\mathcal{P}^2_{\mathrm{MSM}} \;=\; \frac{1}{m}\left\|\Gamma\left(\mathrm{I}-\hat{\mathsf{C}}\right)^{-1}\left(\frac{1}{\|\boldsymbol{\omega}\|_2^2}\hat{\mathsf{C}}\mathsf{R}\boldsymbol{\omega}\boldsymbol{\omega}^\top - \mathsf{R} + \Gamma\left(\hat{\mathsf{C}}\right)\mathsf{R}\left(\mathrm{I}-\frac{1}{\|\boldsymbol{\omega}\|_2^2}\boldsymbol{\omega}\boldsymbol{\omega}^\top\right)\right)\right\|_{\mathcal{F}}^2.$$

The lemma. We want to show that $\bar{\mathbf{x}}_{(j)}$ as defined by equation (9) is given by equation (10). We start by expanding the right-hand side of equation (10) by substituting $\tilde{\mathsf{R}}_j$ from (11), giving:

$$\begin{aligned}\frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger\tilde{\mathsf{R}}_j\boldsymbol{\omega} &= \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger\mathsf{D}_j\mathsf{R}\boldsymbol{\omega} + \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger(\mathrm{I}-\mathsf{D}_j)\mathsf{C}\bar{\mathbf{x}}_{(j)}\boldsymbol{\omega}^\top\boldsymbol{\omega}\\ &= \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger\mathsf{D}_j\mathsf{R}\boldsymbol{\omega} + \mathsf{C}^\dagger\mathsf{C}\bar{\mathbf{x}}_{(j)} - \mathsf{C}^\dagger\mathsf{D}_j\mathsf{C}\bar{\mathbf{x}}_{(j)},\end{aligned}$$

since $\boldsymbol{\omega}^\top\boldsymbol{\omega} = \|\boldsymbol{\omega}\|_2^2$. The second term reduces to $\bar{\mathbf{x}}_{(j)}$ since $\mathsf{C}^\dagger\mathsf{C} = \mathrm{I}$. By replacing $\bar{\mathbf{x}}_{(j)}$ by its expression (9), the third term expands as:

$$\begin{aligned}\mathsf{C}^\dagger\mathsf{D}_j\mathsf{C}\bar{\mathbf{x}}_{(j)} &= \frac{1}{\|\boldsymbol{\omega}\|_2^2}\left(\mathsf{C}^\top\mathsf{C}\right)^{-1}\mathsf{C}^\top\mathsf{D}_j\mathsf{C}\left(\mathsf{C}^\top\mathsf{D}_j\mathsf{C}\right)^{-1}\mathsf{C}^\top\mathsf{D}_j\mathsf{R}\boldsymbol{\omega}\\ &= \frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger\mathsf{D}_j\mathsf{R}\boldsymbol{\omega},\end{aligned}$$

and the overall expression simplifies to:

$$\frac{1}{\|\boldsymbol{\omega}\|_2^2}\mathsf{C}^\dagger\tilde{\mathsf{R}}_j\boldsymbol{\omega} \;=\; \bar{\mathbf{x}}_{(j)}.$$

5 Application to Estimating Rigid Affine Thin-Plate Spline Warps

A warp is a geometric $\mathbb{R}^2 \to \mathbb{R}^2$ transformation matching corresponding pixels between two images. Estimating a warp from point correspondences in two images is one of the most important problems in fields such as computer vision, medical image analysis, photogrammetry and augmented reality. The warp is often parameterized by some smooth transformation driven by deformation centres. This section shows an application of the non-iterative PRESS formula for multiple scaled measurement sets to the problem of assessing the quality of an image warp called the Rigid Affine Thin-Plate Spline Warp (RA-Warp), that we originally proposed in [3]. We show that the number of deformation centres can be selected by minimizing the PRESS, which corresponds to maximizing the predictivity of the warp. The fact that the RA-Warp is an empirical model of the image flow field implies that the fitting residuals

do not follow a simple parametric noise model, which rules out 'classical' model selection techniques such AIC, BIC and MDL [7]. In our previous work [2], we showed how leave-one-out cross-validation (as described in §2 *i.e.*, for a single parameter set and a single measurement set) can be used to select the regularization parameter of a deformable affine warp.

5.1 The RA-Warp and its PRESS

A warp is an $\mathbb{R}^2$ to $\mathbb{R}^2$ function that models the deformation or flow field between two images. The RA-Warp is dedicated to the case of two images of a rigid smooth surface, and models the cameras as affine *i.e.*, parallel, projections. An example is shown in figure 1.

Figure 1: The two example images we use, overlaid with the $m = 70$ point correspondences and corresponding epipolar lines. These two images show a bed sheet wrapped around a chair. They were taken while strongly zooming, making the camera close to affine. The bed sheet remained still between the two snapshots and thus only the relative position and orientation of the camera changed. The epipolar lines depend on this relative camera motion which is computed from the point correspondences.

Let $\mathbf{q} \in \mathbb{R}^2$ be the coordinates of a point in the first image. The RA-Warp depends on a parameter vector $\boldsymbol{\delta} \in \mathbb{R}^l$ and writes as:

$$\mathcal{W}(\mathbf{q};\boldsymbol{\delta}) \stackrel{\text{def}}{=} \mathcal{S}\tilde{\mathbf{q}} + \tau(\mathbf{q};\boldsymbol{\delta})\mathbf{s} \qquad \text{with} \qquad \tilde{\mathbf{q}}^\top = \begin{pmatrix} \mathbf{q}^\top & 1 \end{pmatrix},$$

where τ is an $\mathbb{R}^2$ to $\mathbb{R}$ Thin-Plate Spline as derived by Duchon [5], and $(\mathcal{S};\mathbf{s})$ are camera parameters[4] that we estimate from image point correspondences as detailed in appendix A. Note that $\mathbf{s}$ is the direction of the epipolar lines in

[4]These are related to the parameters of the second camera matrix in the canonical basis of the 3D space for which the first camera matrix is (I $\mathbf{0}$). More specifically, $\mathcal{S}$ is an affine transformation between the two images related to some arbitrary reference plane, and $\mathbf{s}$ is the epipole in the second image, lying at infinity.

the second image. Without loss of generality, we normalize it using $\|\mathbf{s}\|_2^2 = 1$. This warp guarantees that the points move along epipolar lines. Other types of smooth functions could be used for τ, such as a tensor-product of spline functions.

The parameter vector $\boldsymbol{\delta}$ of the Thin-Plate Spline τ contains the depths of some l deformation centres. The Thin-Plate Spline is written $\tau(\mathbf{q};\boldsymbol{\delta}) = \ell_{\mathbf{q}}^\top \mathcal{Y}\boldsymbol{\delta}$, where $\mathcal{Y}$ is a constant matrix depending on the deformation centres in the first image, and $\ell_{\mathbf{q}}$ is a nonlinear lifting function depending on point $\mathbf{q}$. Details on the derivation of τ are given in appendix B.

Given m point correspondences $\mathbf{q}_j \leftrightarrow \mathbf{q}'_j$, the camera parameters $(\mathcal{S};\mathbf{s})$ and the deformation centres in $\mathcal{Y}$, we estimate the parameter vector $\boldsymbol{\delta}$ by minimizing the following cost function, measuring the euclidean distance between the points in the second image and those transferred by the warp from the first image:

$$\mathcal{E}^2_{\mathrm{RA}}(\boldsymbol{\delta}) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{j=1}^{m} \|\mathcal{W}(\mathbf{q}_j,\boldsymbol{\delta}) - \mathbf{q}'_j\|_2^2 = \frac{1}{m}\sum_{j=1}^{m} \|\mathcal{S}\tilde{\mathbf{q}}_j + \tau(\mathbf{q}_j;\boldsymbol{\delta})\mathbf{s} - \mathbf{q}'_j\|_2^2. \tag{14}$$

Substituting the expression of the Thin-Plate Spline τ, we get:

$$\mathcal{E}^2_{\mathrm{RA}}(\boldsymbol{\delta}) = \frac{1}{m}\sum_{j=1}^{m} \left\|\mathcal{S}\tilde{\mathbf{q}}_j + \left(\ell_{\mathbf{q}_j}^\top \mathcal{Y}\boldsymbol{\delta}\right)\mathbf{s} - \mathbf{q}'_j\right\|_2^2.$$

Transposing and gathering all the measurements in matrices $\tilde{\mathcal{Q}}$ and $\mathcal{Q}'$ whose rows are $\tilde{\mathbf{q}}_j$ and $\mathbf{q}'_j$ respectively, and using the vectors $\ell_{\mathbf{q}_j}$ for the rows of matrix $\mathcal{M}$, gives:

$$\mathcal{E}^2_{\mathrm{RA}}(\boldsymbol{\delta}) = \frac{1}{m}\left\|\tilde{\mathcal{Q}}\mathcal{S}^\top + \mathcal{M}\mathcal{Y}\boldsymbol{\delta}\mathbf{s}^\top - \mathcal{Q}'\right\|_{\mathcal{F}}^2.$$

We identify this as a multiple scaled measurement sets linear least squares problem, as defined by equation (4). We get the solution from equation (5) as:

$$\bar{\boldsymbol{\delta}} = (\mathcal{M}\mathcal{Y})^\dagger \left(\mathcal{Q}' - \tilde{\mathcal{Q}}\mathcal{S}^\top\right)\mathbf{s}, \tag{15}$$

since $\|\mathbf{s}\|_2^2 = 1$. Using the non-iterative PRESS formula (7), we obtain the PRESS statistic for the RA-Warp as:

$$\mathcal{P}^2_{\mathrm{RA}} \stackrel{\text{def}}{=} \frac{1}{m}\left\|\Gamma\left(\mathrm{I} - \widehat{\mathcal{M}\mathcal{Y}}\right)^{-1}\left(\widehat{\mathcal{M}\mathcal{Y}}\left(\mathcal{Q}' - \tilde{\mathcal{Q}}\mathcal{S}^\top\right)\mathbf{s}\mathbf{s}^\top - \mathcal{Q}' + \tilde{\mathcal{Q}}\mathcal{S}^\top + \Gamma\left(\widehat{\mathcal{M}\mathcal{Y}}\right)\left(\mathcal{Q}' - \tilde{\mathcal{Q}}\mathcal{S}^\top\right)\left(\mathrm{I} - \mathbf{s}\mathbf{s}^\top\right)\right)\right\|_{\mathcal{F}}^2. \tag{16}$$

5.2 Estimation Algorithm

In practice, we are given a set of point correspondences from which we can estimate the camera parameters $(\mathcal{S};\mathbf{s})$. We do not however know how many deformation centres l are needed, and where to place them in the first image. A sensible choice, though heuristic, is to choose as deformation centres the vertices of a regular, square grid located in the vicinity of the data points. Choosing the number of deformation centres is more critical: underestimating l

makes the warp too constrained to explain the measurements, while overestimating l makes it too flexible, possibly ill-conditioned, with bad predictivity. We propose to use the PRESS statistic in order to choose the number of deformation centres. We start with a coarse square control grid of, say, $l_{\min} = 2^2 = 4$ deformation centres, and subdivide it while monitoring the PRESS, until the maximum possible number of deformation centres $l_{\max} = \lfloor \sqrt{m-1} \rfloor^2$ is reached (which guarantees that the design matrix in the linear least squares systems is not rank deficient). Note that there is no guarantee for the PRESS to be a convex function of l, even though it very often is, in practice.

5.3 Experimental Results

For the example data shown in figure 1, we have $m = 70$ point correspondences. We thus try to fit the warp driven by grids of $l_{\min} = 4, 9, 16, 25, 36, 49, 64 = l_{\max}$ deformation centres. The fitting Root Mean Square Residual (RMSR) and PRESS as functions of the number of deformation centres[5] are shown in figure 2. They were computed using all the 70 point correspondences.

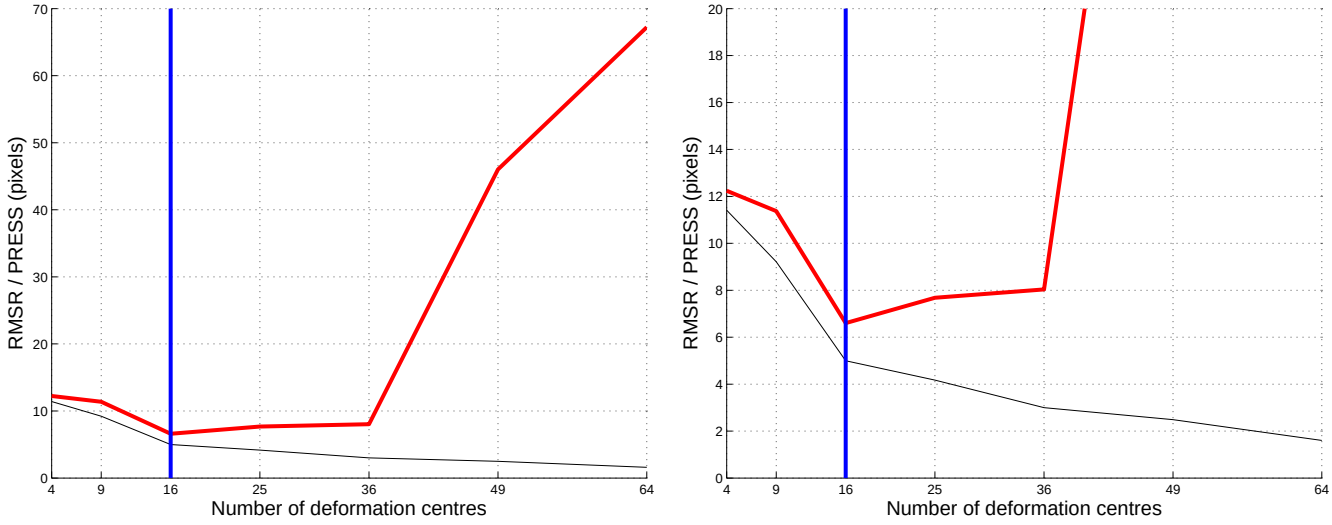
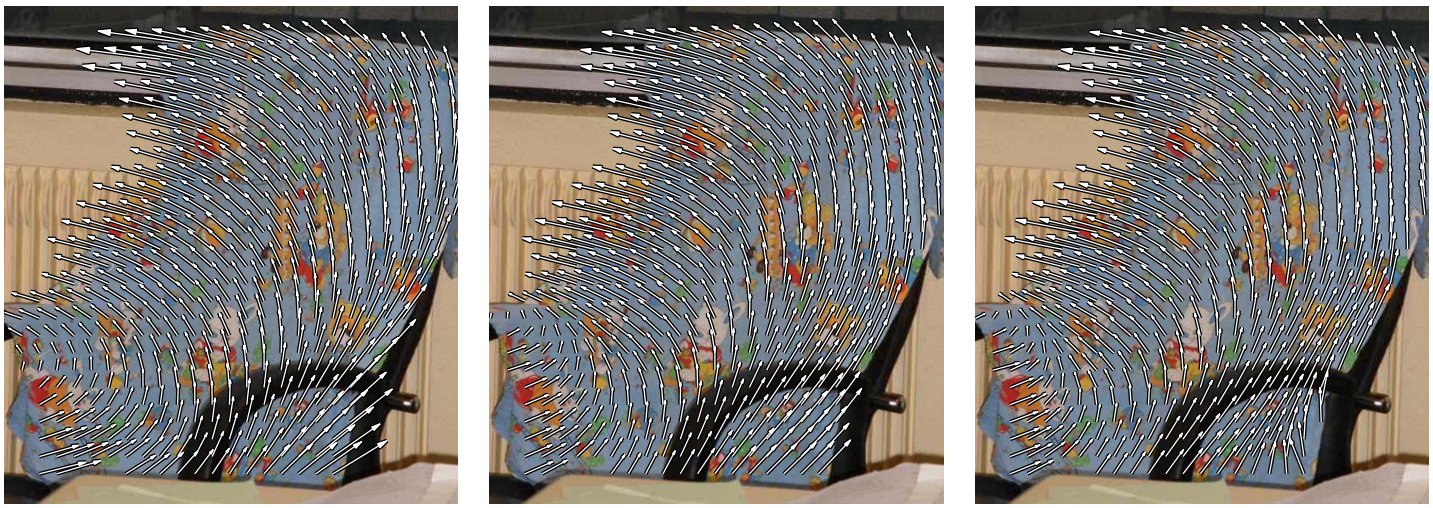


Figure 2: The fitting residual (RMSR, thin curve) and prediction sum of squares (PRESS, thick curve) as functions of the number of deformation centres. The vertical line shows the selected number of deformation centres corresponding to the minimum PRESS. The two graphs are similar: the right one is a zoom on the left one.

We observe that the residual RMSR decreases as the number of deformation centres increases. This is to be expected since the most parameters in the model the least the fitting error. The PRESS decreases until 16 deformation centres are reached, and then grows as the number of deformation centres increases beyond 16. This is explained as follows: for less than 16 deformation centres, the warp is not flexible enough to model the actual image deformations, while for more than 16 deformation centres, the warp is too flexible and is less and less well constrained by the point correspondences. In both cases, it fails to accurately capture the deformation, and thus does not interpolate well the data leading to a bad predictivity.

[5]The RMSR $\mathcal{E}_{\mathrm{RA}}(\bar{\boldsymbol{\delta}})$ is given by evaluating the transfer error of equation (14) at the optimal solution $\bar{\boldsymbol{\delta}}$ given by equation (15). The PRESS $\mathcal{P}_{\mathrm{RA}}$ is given by equation (16).

The so-called flow field is the set of displacement vectors for each pixel in one of the two images. The flow field sub-sampled to $\frac{1}{20^2}$ pixels is shown in figure 3 for different numbers of deformation centres. Noticeable differences can



9 deformation centres | 16 deformation centres | 25 deformation centres

Figure 3: The estimated flow field for different numbers l of deformation centres. The solution minimizing the PRESS is with $l = 16$ deformation centres.

be seen, though it is difficult to visually figure out which solution is the best one. This is better seen by observing the color discrepancy image, computed by warping the second image onto the first one and taking the absolute value of their difference. This is shown in figure 4 over a region of interest defined as the area covered by the bed sheet in the source image. The color discrepancy, computed as the Root Mean Square of the color discrepancy image, is

9 deformation centres
color discrepancy: 67.02

16 deformation centres
color discrepancy: 59.90

25 deformation centres
color discrepancy: 63.84

Figure 4: The color discrepancy images for different numbers l of deformation centres. Black indicates low discrepancy, and white indicates high discrepancy. The solution minimizing the PRESS is with $l = 16$ deformation centres.

lower for 16 deformation centres than for 9 and 25. It also is visually seen that for 16 deformation centres, only the area under the arm of the chair shows high discrepancy, while the rest of the bed sheet area is correctly registered,

which is not the case for 9 and 25 deformation centres.

Finally, we report an experiment illustrating to which extent the PRESS statistic is influenced by the number of data points. We drew random subsets of $m' = 5, \ldots, 70$ data points. For each subset, we computed the 'best' number of deformation centres, chosen in $\{4, 9, \ldots, \lfloor\sqrt{m'-1}\rfloor^2\}$. Average values over 1,000 different subsets for each value of m' were computed. They are plotted on the graph shown in figure 5, along with the standard deviation.

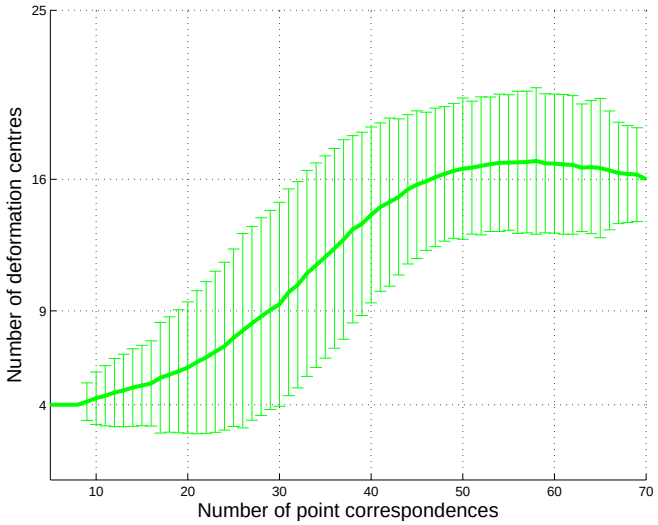


Figure 5: This graph shows the average and standard deviation of the number of selected deformation centres based on the prediction sum of squares (PRESS) as a function of the number of data points.

We observe that as the number of data points increases, the number of selected deformation centres increases. The slight decrease as $m' \to m$ is due to the fact that the total number of data points is approached. What can be first noticed is that for $m' \geq 40$ the selected number of deformation centres is consistent, always being $l = 16$. For $16 < m' < 40$, the selected number of deformation centres grows from 4 to 9 ($m' > 16$ is a necessary condition to select 16 deformation centres) For $4 < m' \leq 16$, we consistently get $l = 4$ deformation centres. This makes sense since it is difficult to infer the full image warp from such a limited amount of data, and we thus expect the selected warp to have a low complexity. Overall, this result shows that the PRESS selected number of centres is quite robust to variations in the number of data points, since the same result is consistently obtained for $40 \leq m' \leq 70$ data points.

6 Conclusion

We derived non-iterative formulae for the Prediction Sum of Squares (PRESS) statistic for linear least squares with multiple parameter or measurement sets. As an application, we showed that this can be used to assess the quality of an image warp. It also allows one to select its complexity by varying the number of deformation centres and shows to be consistent with the color discrepancy image.

There are some open research directions with the formulae we propose. The first one is to extend them to the case of restricted least squares, as was done in [11] for standard linear least squares. The second one is to investigate

how they can be used to select a regularization parameter for damped least squares, by replacing the hat matrix by the influence matrix. Finally, we assumed for the warp estimation application that the deformation centres were placed on a regular grid. A 'stochastic deformation centre placement' procedure could be imagined, whereby one would draw random sets of deformation centres, and keep the one minimizing the PRESS statistic.

A Computing the Camera Parameters

The camera parameters are extracted from the affine fundamental matrix $\mathcal{A}$ that can easily be estimated from point correspondences, and which represents an implicit reconstruction of the two cameras. The affine fundamental matrix is partitioned as:

$$\mathcal{A} \stackrel{\text{def}}{=} \begin{pmatrix} 0 & 0 & \\ 0 & 0 & \boldsymbol{\imath} \\ & \boldsymbol{\jmath}^\top & \end{pmatrix} \quad \text{with} \quad \boldsymbol{\jmath}^\top \stackrel{\text{def}}{=} (c \;\; d \;\; e) \quad \text{and} \quad \boldsymbol{\imath}^\top \stackrel{\text{def}}{=} (a \;\; b). \tag{17}$$

Estimating the affine fundamental matrix. We give a simple procedure for estimating the affine fundamental matrix by minimizing the transfer error *i.e.*, the sum of squared distances between the epipolar lines and the data points in the second image. Even though there exist a maximum likelihood algorithm as described in *e.g.* [8], we prefer to minimize the transfer error for the sake of consistency with the cost function (14) used next to estimate the RA-Warp. The problem can be written:

$$\min_{\mathcal{A}} \sum_{j=1}^{m} d_\perp^2(\mathbf{q}'_j, \mathcal{A}\tilde{\mathbf{q}}_j).$$

Expanding the point to line orthogonal distance $d_\perp$ by its expression and replacing $\mathcal{A}$ by its expression (17), we get:

$$\min_{\boldsymbol{\imath},\boldsymbol{\jmath}} \frac{1}{\|\boldsymbol{\imath}\|^2} \sum_{j=1}^{m} (\boldsymbol{\jmath}^\top \tilde{\mathbf{q}}_j + \boldsymbol{\imath}^\top \mathbf{q}'_j)^2.$$

This is not the most common minimization problems one meets. The difficulty stems from the leading factor. In order to get rid of it, we fix the arbitrary scale of the fundamental matrix using $\|\boldsymbol{\imath}\|^2 = 1$. This makes sense since $\|\boldsymbol{\imath}\|^2$ cannot vanish for otherwise $\mathcal{A}$ would become rank one. The problem turns into a constrained homogeneous linear least squares minimization:

$$\min_{\boldsymbol{\imath},\boldsymbol{\jmath}} \sum_{j=1}^{m} (\boldsymbol{\jmath}^\top \tilde{\mathbf{q}}_j + \boldsymbol{\imath}^\top \mathbf{q}'_j)^2 \qquad \text{s.t.} \qquad \|\boldsymbol{\imath}\|^2 = 1.$$

We rewrite the problem in matrix form:

$$\min_{\boldsymbol{\imath},\boldsymbol{\jmath}} \|\mathcal{Q}'\boldsymbol{\imath} + \tilde{\mathcal{Q}}\boldsymbol{\jmath}\|^2 \qquad \text{s.t.} \qquad \|\boldsymbol{\imath}\|^2 = 1,$$

where the rows of matrices $\tilde{\mathcal{Q}}$ and $\mathcal{Q}'$ are $\tilde{\mathbf{q}}_j$ and $\mathbf{q}'_j$ respectively. Setting $\boldsymbol{\jmath} = -\tilde{\mathcal{Q}}^\dagger \mathcal{Q}' \boldsymbol{\imath}$, substituting in the previous equation and factorizing gives:

$$\min_{\boldsymbol{\imath}} \|(\mathrm{I} - \tilde{\mathcal{Q}}\tilde{\mathcal{Q}}^\dagger)\mathcal{Q}'\boldsymbol{\imath}\|^2 \qquad \text{s.t.} \qquad \|\boldsymbol{\imath}\|^2 = 1,$$

that we solve using the standard method based on the Singular Value Decomposition, as described for instance in [8, A5.3].

Extracting the camera parameters. We follow the canonical realization framework of [9] for perspective cameras. The canonical camera for the first image is $(\mathrm{I} \ \mathbf{0})$. The canonical camera for the second image is $([\tilde{\mathbf{e}}']_\times \mathcal{A} \ \tilde{\mathbf{e}}')$ where $\tilde{\mathbf{e}}'$ is the epipole in the second image. In the affine case, $\tilde{\mathbf{e}}'^\top \stackrel{\text{def}}{=} (-b \ a \ 0)$, and we observe that $\|\tilde{\mathbf{e}}'\| = \|\boldsymbol{\imath}\| = 1$. The camera parameters $(\mathcal{S}; \mathbf{s})$ for the second image are thus obtained as:

$$\mathcal{S} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} [\tilde{\mathbf{e}}']_\times \mathcal{A} = \boldsymbol{\imath}\boldsymbol{\jmath}^\top \qquad \text{and} \qquad \mathbf{s} \stackrel{\text{def}}{=} \mathbf{e}' = (-b \ a)^\top.$$

B The Thin-Plate Spline

The Thin-Plate Spline is an $\mathbb{R}^2 \to \mathbb{R}$ function driven by assigning target values α_k to 2D deformation centres $\mathbf{p}_k$ and enforcing several conditions: the Thin-Plate Spline is the Radial Basis Function (RBF) that minimizes the integral bending energy. The idea of using the Thin-Plate equation as an interpolation map is due to Duchon [5]. Note that the standard $\mathbb{R}^2 \to \mathbb{R}^2$ Thin-Plate Spline Warp of [4] (or Deformable Affine Thin-Plate Spline Warp with our naming conventions) is obtained by stacking two Thin-Plate Splines sharing their deformation centres.

Standard parameterization. Given the l deformation centres $\mathbf{p}_k$ in the first image, the Thin-Plate Spline is usually parameterized by an $l+3$ coefficient vector $\mathbf{h}^\top = (\ \mathbf{w}^\top \quad \mathbf{f}^\top)$ and a regularization parameter $\lambda \in \mathbb{R}^+$. These coefficients can be computed from the target vector $\boldsymbol{\alpha}$. There are l coefficients in $\mathbf{w}$ and three coefficients in $\mathbf{f}$. The Thin-Plate Spline is given by:

$$\eta(\mathbf{q}; \mathbf{h}) \stackrel{\text{def}}{=} \left(\sum_{k=1}^{l} w_k \ \rho(d^2(\mathbf{q}, \mathbf{p}_k)) \right) + \mathbf{f}^\top \tilde{\mathbf{q}}, \tag{18}$$

where $\rho(d) = d \log(d)$ is the Thin-Plate Spline kernel function for the squared distance. The coefficients in $\mathbf{w}$ must satisfie $\tilde{\mathcal{P}}^\top \mathbf{w} = \mathbf{0}$, where the rows of $\tilde{\mathcal{P}}$ are the $\tilde{\mathbf{p}}_k$. These three 'side-conditions' ensure that the Thin-Plate Spline has square integrable second derivatives. It is convenient to define the $(l+3)$-vector $\ell_{\mathbf{q}}$ as:

$$\ell_{\mathbf{q}}^\top \stackrel{\text{def}}{=} (\rho(d^2(\mathbf{q}, \mathbf{p}_1)) \ \cdots \ \rho(d^2(\mathbf{q}, \mathbf{p}_l)) \ \tilde{\mathbf{q}}^\top),$$

allowing the Thin-Plate Spline (18) to be rewritten as a dot product:

$$\eta(\mathbf{q};\mathbf{h}) \quad = \quad \ell_{\mathbf{q}}^{\top}\mathbf{h}. \tag{19}$$

Interpolating the α_k. Applying the Thin-Plate Spline (18) to the centre $\mathbf{p}_r$ with target value α_r gives:

$$\left(\sum_{k=1}^{l} w_k \ \rho(d^2(\mathbf{p}_r,\mathbf{p}_k))\right) + \mathbf{f}^{\top}\tilde{\mathbf{p}}_r \quad = \quad \alpha_r.$$

Combining the equations obtained for all the l centres with the side-conditions $\tilde{\mathcal{P}}^{\top}\mathbf{w} = \mathbf{0}$ in a single matrix equation gives:

$$\underbrace{\begin{pmatrix} \mathsf{K} & \tilde{\mathcal{P}} \\ \tilde{\mathcal{P}}^{\top} & 0 \end{pmatrix}}_{\mathcal{D}} \underbrace{\begin{pmatrix} \mathbf{w} \\ \mathbf{f} \end{pmatrix}}_{\mathbf{h}} = \begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix} \qquad \text{with} \qquad K_{r,k} \ = \ \begin{cases} \lambda & r = k \\ \rho(d^2(\mathbf{p}_r,\mathbf{p}_k)) & \text{otherwise} \end{cases}$$

Adding λI to the leading block K of the design matrix $\mathcal{D}$ acts as a regularizer. Solving for $\mathbf{h}$ by inverting $\mathcal{D}$ is the classical linear method for estimating the Thin-Plate Spline coefficients due to Bookstein [4]. The coefficient vector $\mathbf{h}$ is thus a nonlinear function of the regularization parameter λ and a linear function of the target vector $\boldsymbol{\alpha}$.

A feature-driven parameterization. We express $\mathbf{h}$ as a linear 'back-projection' of the target vector $\boldsymbol{\alpha}$. This is modeled by the matrix $\mathcal{Y}$, nonlinearly depending on λ, given by the l leading columns of $\mathcal{D}^{-1}$:

$$\mathbf{h} \quad = \quad \mathcal{Y}\boldsymbol{\alpha} \qquad \text{with} \qquad \mathcal{Y} \ \stackrel{\text{def}}{=} \ \begin{pmatrix} \mathsf{K}^{-1}\left(\mathrm{I} - \tilde{\mathcal{P}}\left(\tilde{\mathcal{P}}^{\top}\mathsf{K}^{-1}\tilde{\mathcal{P}}\right)^{-1}\tilde{\mathcal{P}}^{\top}\mathsf{K}^{-1}\right) \\ \left(\tilde{\mathcal{P}}^{\top}\mathsf{K}^{-1}\tilde{\mathcal{P}}\right)^{-1}\tilde{\mathcal{P}}^{\top}\mathsf{K}^{-1} \end{pmatrix}. \tag{20}$$

This parameterization has the advantages to separate λ and $\boldsymbol{\alpha}$ and introduces units[6]. The side-conditions are naturally enforced by this parameterization.

Incorporating the parameterization (20) into the Thin-Plate Spline (19) we obtain what we call the *feature-driven* parameterization $\tau(\mathbf{q};\boldsymbol{\alpha}) = \eta(\mathbf{q};\mathbf{h})$ for the Thin-Plate Spline:

$$\tau(\mathbf{q};\boldsymbol{\alpha}) \ \stackrel{\text{def}}{=} \ \ell_{\mathbf{q}}^{\top}\mathcal{Y}\boldsymbol{\alpha}.$$

[6] While $\mathbf{h}$ has no obvious unit, $\boldsymbol{\alpha}$ in general has (*e.g.* pixels, meters).

References

[1] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[2] 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.

[3] A. Bartoli, M. Perriollat, and S. Chambon. Generalized Thin-Plate Spline warps. In *International Conference on Computer Vision and Pattern Recognition*, 2007.

[4] 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.

[5] J. Duchon. Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *RAIRO Analyse Numérique*, 10:5–12, 1976.

[6] J. E. Gentle, W. Hardle, and Y. Mori. *Handbook of Computational Statistics*. Springer-Verlag, 2004.

[7] N. Gheissari and A. Bab-Hadiashar. A comparative study of model selection criteria for computer vision applications. *Image and Vision Computing*, 26:1636–1649, 2008.

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

[9] Q.-T. Luong and T. Vieville. Canonic representations for the geometries of multiple projective views. *Computer Vision and Image Understanding*, 64(2):193–229, 1996.

[10] D. Montgomery and E. Peck. *Introduction to Linear Regression Analysis*. Wiley, New York, USA, 1992.

[11] T. Tarpey. A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician*, 54(2):116–118, May 2000.

[12] G. Wahba and S. Wold. A completely automatic French curve: Fitting spline functions by cross-validation. *Communications in Statistics*, 4:1–17, 1975.