

# Imputation on 2-dimensional data via lifting

Robert G. Aykroyd, Stuart Barber & Samuel J. Peck\*

Department of Statistics, University of Leeds

## 1 Introduction and setting

We describe a method for imputing from a grid of irregularly spaced data points on to any other grid using a Voronoi based lifting scheme. The lifting scheme is a generalisation of wavelet decompositions. Here, the lifting scheme is used as an interpolating/smoothing method for data on an irregular grid. We show an example of this method on real data, and a comparison to similar methods – Heaton and Silverman’s (2008) imputation method and Kriging.

Suppose we have some data collected on an irregular two-dimensional grid. Let  $X_F = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$  be the matrix of data points, where  $\mathbf{x}_i$  is the location of the  $i^{\text{th}}$  data point. Also let  $f_i = f(\mathbf{x}_i), i = 1, \dots, n$  be function values at the data points observed with error, i.e.  $f(\mathbf{x}) = g(\mathbf{x}) + \epsilon$  where  $g(\cdot)$  is the true function and  $\epsilon \sim N(0, \sigma^2)$ .

We wish to make estimates of the function value at points where we have no observations, say  $X_M = (\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+m})^T$ . To make these estimates, we perform a lifting transformation on the combined data sets and estimate the missing values  $f_i, i = n + 1, \dots, n + m$  based on the expected sparsity of lifting coefficients.

## 2 Lifting

We combine the fixed and missing data as  $X = \begin{bmatrix} X_F \\ X_M \end{bmatrix}$ . The data are decomposed using the leave-one-out Voronoi scheme from Jansen *et al.* (2004a), yielding a transformation matrix  $W$  of dimension  $(n+m) \times (n+m)$ . Naturally, the order in which the points are lifted is important, and this is particularly true of the missing points. However, we will not address this here as, although it will affect the results, it does not change the calculations.

## 3 Imputation method

We have known function values  $f_1, \dots, f_n$  for the fixed points  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , and unknown function values  $f_{n+1}, \dots, f_{n+m}$  corresponding to the missing points  $\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+m}$ . It is these unknown function values that we want to estimate.

This particular lifting scheme differs from the classical wavelet transform in that, while wavelet coefficients generally describe local differences between 2 or more function values, lifting coefficients (in this scheme) describe the difference between a *particular* point and its neighbours. This means that there is a 1-1 correspondence between the original data points (or their function values) and the lifting coefficients – the lifting coefficient for a point is created when that point is lifted. The imputation method works by making assumptions about the lifting coefficients corresponding to the missing points and estimates of the function values at these points can be made based on those assumptions.

Let  $\mathbf{f} = (f_1, \dots, f_{n+m})$ , and let  $\mathbf{z} = W\mathbf{f}$ . This is possible because the construction of  $W$  involves only the locations  $X$ , and not the (partially missing) function values  $\mathbf{f}$ . Here,  $\mathbf{z}$  is the vector of lifting coefficients. Naturally,  $z_i$  is the lifting coefficient corresponding to  $\mathbf{x}_i$  and  $f_i$ . We need to make assumptions about the coefficients corresponding to missing points, i.e.  $z_{n+1}, \dots, z_{n+m}$ . From now on, we assume that these values are fixed;  $z_{n+i} = \zeta_i$  where  $\{\zeta_i : i = 1, \dots, m\}$  are constants. For instance, since we expect our coefficients to be “sparse” – that is, the majority close to zero – a reasonable choice would be  $\zeta_i = 0$  for all  $i = 1, \dots, m$ .

## 4 Derivation

Here, we will look at the simplest case – that where we have just one missing value;  $m = 1$ . Hence  $X_M = \mathbf{x}_{n+1}^T$  and our unknown is just the single value  $f_{n+1}$ . Also, let  $W^* = (w_{ij}^*) = W^{-1}$ ; then  $\mathbf{f} = W^* \mathbf{z}$ . We derive our estimate of this value as follows. The function value at the missing point has the form

$$f_{n+1} = \sum_{i=1}^{n+1} w_{n+1,i}^* z_i = \sum_{i=1}^n w_{n+1,i}^* z_i + w_{n+1,n+1}^* z_{n+1},$$

so our estimate is

$$\hat{f}_{n+1} = \sum_{i=1}^n w_{n+1,i}^* z_i + w_{n+1,n+1}^* \zeta_1.$$

The  $z_i$  also depend on  $f_{n+1}$ , since

$$z_i = \sum_{j=1}^{n+1} w_{ij} f_j = \sum_{j=1}^n w_{ij} f_j + w_{i,n+1} f_{n+1},$$

for  $i = 1, \dots, n$ . Substitution gives

$$\hat{f}_{n+1} = \sum_{i=1}^n \sum_{j=1}^n w_{n+1,i}^* w_{ij} f_j + \hat{f}_{n+1} \sum_{i=1}^n w_{n+1,i}^* w_{i,n+1} + w_{n+1,n+1}^* \zeta_1.$$

By re-arranging, we find the estimate is

$$\hat{f}_{n+1} = \frac{1}{1 - \sum_{i=1}^n w_{n+1,i}^* w_{i,n+1}} \left( \sum_{i=1}^n \sum_{j=1}^n w_{n+1,i}^* w_{ij} f_j + w_{n+1,n+1}^* \zeta_1 \right).$$

## 5 Results

### 5.1 Real data

We apply our imputation method to a real data set of pest incidences at selected locations in England and Wales.

In figure 5.1, the left plot shows the observed data with the intensity of infestation shown by the colour and size of the dot – white is no infestation, black is heavy infestation. The right plot shows the imputed surface obtained using our method, a portion of a  $100 \times 100$  uniform grid. The intensity scale is the same in both plots. We see that the method picks up the high intensity region in the east, as well as the region of very low intensity in the south. Also, in areas where there is no data, the estimates are almost constant.



Figure 1: Real data example: Pest infestation in UK. Left plot: Observation locations with intensity of infestation. Right plot: Imputed surface using our lifting imputation method.

## 5.2 Simulated example

We compare our imputation to Heaton and Silverman’s MCMC based approach and Kriging using a simulated example. We apply the methods to a number of test functions – namely, the two-dimensional analogues of the standard test functions used introduced in Donoho and Johnstone (1994): Bumps, Blocks, Heavisine and Doppler, and “maartenfunc”. All of these functions are defined in Jansen *et al.* (2004b).

To save computational time, we set observed points to be on the same grid as the real data (re-scaled so that the test functions and data points are on the same scale), and impute on to a selection of 10 points from the original  $100 \times 100$  grid. This allows us to re-use the transformation matrices generated for the real data. For our method we have two versions: one with smoothing on whitened coefficients performed by Ebayesthresh (see Johnstone and Silverman (2005)) with default settings, and the other on the raw data (no smoothing).

The methods are tested using 100 noisy realisations of the observed data from the test functions. The mean square errors are calculated for each run, and an average of these is taken for our results. These are presented below:

| Test function | HS    |         | K            |         | ABP (S)      |         | ABP (NS)     |         |
|---------------|-------|---------|--------------|---------|--------------|---------|--------------|---------|
| Doppler       | 0.791 | (0.993) | 0.105        | (0.023) | 0.193        | (0.121) | <b>0.079</b> | (0.019) |
| Heavisine     | 0.456 | (0.048) | 0.635        | (0.072) | <b>0.350</b> | (0.057) | 0.559        | (0.067) |
| Blocks        | 70.64 | (33.40) | 0.660        | (0.223) | 3.252        | (0.537) | <b>0.425</b> | (0.179) |
| Bumps         | 0.664 | (0.458) | 2.277        | (0.431) | 3.129        | (0.590) | <b>0.214</b> | (0.113) |
| Maartenfunc   | 1.744 | (1.506) | <b>0.476</b> | (0.196) | 1.025        | (0.847) | 0.537        | (0.186) |

Table 1: Comparison of imputation methods on test functions. Key: HS = Heaton-Silverman method, K = Kriging, ABP (S) = our method with smoothing before imputation, ABP (NS) = our method without smoothing. Values displayed are medians of mean square errors, with the median absolute deviation of these in brackets.

The computational time required for each method should also be mentioned. Since the Heaton-Silverman method uses a Gibbs sampler to obtain its estimates, it is very computationally intensive. Our method, like Kriging, is a deterministic method, whose results are calculated almost instantaneously. The only

significant overhead in our method is in computing the lifting decompositions for the missing points. Of course, this overhead is also present in the Heaton-Silverman method, therefore we consider our method to be reasonably computationally efficient.

Table 1 shows that our method is competitive with Kriging, a standard method. In fact, in this example, the MMSE for our method is lower than for Kriging for all test functions except Maartenfunc. Our method is also competitive with Heaton-Silverman, which has bimodality issues, particularly with Blocks; see Heaton and Silverman (2008). Interestingly, our method is better when no smoothing is applied to the data. This may be due to oversmoothing in the Ebayesthresh method on the whitened coefficients, though further investigation is required.

## 6 Future work

The method as it stands produces a point estimate at the missing location. We would like to extend it so that we can produce confidence intervals or densities for our estimate, as the Heaton-Silverman method does. It is also desirable to have some smoothing component in the method, although the smoothing we tried here appeared to hinder, rather than help, the method. Finally, to further test our method, we would like to perform a more complete simulation study, using multiple data grids as well as multiple realisations.

## Acknowledgements

We would like to thank Alistair Murray and Phil Northing (Central Science Laboratory) for providing data, and Central Science Laboratory, EPSRC and the Nuffield Foundation for their funding of this work.

## References

- Donoho, D.L. and Johnstone, I.M. (2004) Ideal spatial adaptation by wavelet shrinkage *Biometrika*, **81**(3), 425–455
- Heaton, T.J., Silverman, B.W. (2008) A wavelet- or lifting-scheme-based imputation method *Journal of the Royal Statistical Society: Series B*, **70**(3), 567–587
- Jansen, M., Nason, G.P. and Silverman, B.W. (2004a) Multiscale methods for graphs and irregular multidimensional data.  
*URL: <http://www.stats.ox.ac.uk/~silverma/pdf/jansennasonsilverman.pdf>*
- Jansen, M., Nason, G.P. and Silverman, B.W. (2004b) Simulations and examples of multivariate non-parametric regression using lifting.  
*URL: <http://www.stats.bris.ac.uk/research/stats/reports/2004/0418.pdf>*
- Johnstone, I.M. and Silverman, B.W. (2005) EbayesThresh: R Programs for Empirical Bayes Thresholding *Journal of Statistical Software*, **12**(8), 1–38