Transformation of measurement uncertainties into low-dimensional feature vector space

A. Alexiadis, S. Ferson, E. A. Patterson
a.alexiadis@liverpool.ac.uk

Abstract

Advances in technology allow the acquisition of data with high spatial and temporal resolution. These datasets are usually accompanied by estimates of the measurement uncertainty, which may be spatially or temporally varying and should be taken into consideration when making decisions based on the data. At the same time, various transformations are commonly implemented to reduce the dimensionality of the datasets for post-processing, or to extract significant features. However, the corresponding uncertainty is not usually represented in the low-dimensional or feature vector space. A method is proposed that maps the measurement uncertainty into the equivalent low-dimensional space with the aid of approximate Bayesian computation, resulting in a distribution that can be used to make statistical inferences. The method involves no assumptions about the probability distribution of the measurement error and is independent of the feature extraction process as demonstrated in three examples. In the first two examples Chebyshev polynomials were used to analyse structural displacements and soil moisture measurements; while in the third, principal component analysis was used to decompose global ocean temperature data. The uses of the method range from supporting decision making in model validation or confirmation, model updating or calibration and tracking changes in condition, such as the characterisation of the El Niño Southern Oscillation.

Usage Notes:

The source code and datasets used to produce the results in the paper are included in two folders. The one labelled Chebyshev_based uses 2D Chebyshev polynomials to decompose spatial datasets and includes the I-beam and soil-moisture data. The folder named PCA_based uses principal component analysis (PCA) to decompose global temperature oceanographic data. MATLAB is needed to load the data and run the code. In order to comply with the CC0 license, **some files needed to execute the algorithm must be downloaded directly from Umberto Picchini's github repository (https://github.com/umbertopicchini/abc_g-and-k).** These files are: **param_mask.m** , **param_unmask.m** and **cov_update.m**. Download and include them in the same directory along the rest of the files.

Content

The root scripts needed to reproduce the results of the paper are the: gk_EXPERIMENTAL_DATA.m and gk_EXPERIMENTAL_DATA_PCA.m.

• **gk_EXPERIMENTAL_DATA.m** the main script where all the following functions connect onto.
• **gk_EXPERIMENTAL_DATA_RUN_modelsimulate.m** reconstructs the drawn feature vector during the approximate Bayesian computation into a spatial dataset.

• **gk_EXPERIMENTAL_DATA_RUN_prior.m** returns the product of independent priors for the feature vector.

• **gk_EXPERIMENTAL_DATA_RUN_summaries_v2**: calculation of pixel-wise differences.

• **cov_update** updating of the covariance matrix.

• **abcmcmc_ABC_EXPERIMENTAL_V3**: the ABC-MCMC algorithm that draws samples from the posterior distribution.

• **newTchebDecomp** decomposes the spatial data into a feature vector using Chebyshev polynomials.

• **newTchebRecon** reconstructs a feature vector into its spatial form using Chebyshev polynomials.

• **newTchebRecon_MODIFIED** same as the previous with minor modifications added to speed up the process of the Chebyshev polynomial generation during ABC.

• **Reconstruction_Using_FULL_Kernels** reconstructs a feature vector into its spatial form using Chebyshev polynomials and provides a series of plots to help identify the quality of decomposition.

• **Reconstruction_Using_Less_Kernels** reconstructs a feature vector into its spatial form using a subset of the initial kernels used for its decompositions

• **tm_kernel_weighting** assists in the selection of the important kernels during decomposition so that the dimensionality of the spatial data is decreased.

The corresponding algorithms in the PCA_based folder perform the same tasks.

Settings

For the datasets where **Chebyshev polynomials** are used the choices regarding the execution of the algorithms are made in the script named **gk_EXPERIMENTAL_DATA.m**

Dataset loading

One can select which dataset to load by uncommenting the corresponding line. For the I-Beam case this is done by uncommenting one of lines 22-23 and then establishing that the correct measurement uncertainty (spatially constant) is defined in line 42 by assigning the value to the **meas_unc_mean** variable.
For the case of spatially varying measurement uncertainty the user must establish that the uncertainty field has the same size as the measured one. For the case of the soil moisture data one can uncomment lines 50-55 to reproduce the results.

Selection of no. of coefficients/kernels

This is done in line 87. The selection should be made following the guidelines outlined in the paper.

Selection of number of evaluations

This is done in line 144 in variable R_mcmc. This selection is a function of the dimensionality of the problem i.e. the higher the number of kernels used to decompose the dataset more evaluations should take place. A good starting value is 30000 for relatively low dimensional problems (up to 9-10 kernels). This could go up to hundreds of thousands in high dimensions (more than 100).

Output (for the case of Chebyshev polynomials)

It should be noted that the resulting posterior samples saved in the ABCMCMC_pilot variable are sorted in terms of significance. This means that if the fifth shape descriptor is the one with the largest magnitude in the measurement's feature vector then it will be in the first column. The order of the shape descriptors can be shown using the command *positions(tm_ind_less_kernel)* .

Output (for the case of PCA)

The coefficients of the components are ordered in descending order. This means that the coefficients corresponding to the first PC are in the first column, the coefficients corresponding to the second PC in the second column and so on.

Visualizing the output

This could be done using the command *corrplot(ABCMCMC_pilot)* – Beware of memory limitations for cases with a big number of kernels. The practitioner should be aware of the initial transitory effects during which the MCMC algorithm adapts its covariance matrix to become more efficient. This means that certain initial evaluations during the period known as the burn-in should be excluded from the final distribution. These should normally be less than 5000-10000 evaluations depending on the behaviour of the algorithm and the dimensionality of the data.

Comparison with model predictions

Model predictions for the I-beam displacement data can be found in the Chebyshev_based folder (FE_UY_DISPL_ROI_1.mat and FE_UY_DISPL_ROI_2.mat) for the regions of interest at the middle and the side of the I-beam respectively. A visual comparison with the outputs of the ABC results is possible only when the same set of coefficients is used during the decomposition of the measured and the predicted fields.

The sources of the data are:

Oceanographic data

Gaillard, F. (2015). ISAS-13 temperature and salinity gridded fields. SEANOE. https://doi. org/10.17882/45945 . – **Open Access**.

The *TEMP-DATA-UNCERTAINTY-DEPTH-10M-4-ROW.mat* file provided in the PCA_based folder ensures that all the data needed to reproduce the results of the paper are in one place. Moreover, it removes the burden of loading and pre-processing the data into a format that can be used for further analysis (the pre-processing steps have been described in the paper). It is important to point out that the 132 temperature and uncertainty fields inside the file correspond to an ocean depth of 10m.

I-beam displacement data

Lampeas G., Pasialis V., Lin X., and Patterson E. A. (2015) 'On the validation of solid mechanics models using optical measurements and data decomposition'. Simulation Modelling Practice and Theory, 52: pp. 92-107 – **Authors have agreed to share the data.**

Both the measurements (*DIC_UY_DISPL_ROI_1.mat*, *DIC_UY_DISPL_ROI_2.mat*) and the finite element predictions (*FE_UY_DISPL_ROI_1.mat*, *FE_UY_DISPL_ROI_2.m*at) for the regions of interest at the middle and the side of the I-beam have been generously provided by Lampeas et al.

Soil moisture data

Kang, J., Jin, R., Li, X. and Zhang Y. (2017). 'Block kriging with measurement errors: A case study of the spatial prediction of soil moisture in the middle reaches of Heihe River Basin'. IEEE Geosci. Remote Sens. Lett. 14: pp. 87–91. **– The soil moisture data used in the paper are the result of digitization of figure 5 (right side BKHEME) in Kang et. al.**

The file containing the soil moisture data is *SOIL_MOISTURE_DATA.mat*.