Reconstruction of a signal from time-averaged data

Alistair B Forbes[1,*]
[1]National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK

Abstract. In measurement applications such as air quality monitoring, sensors record the average of a signal of over a fixed time period, for example, every hour or every day. Sensor networks may involve a number of different types of sensors with different temporal resolutions, for instance, expensive reference sensors recording accurate, hourly averages and inexpensive sensors recording noisy averages every ten minutes. In addition, readings for some time periods may be missing. In this paper, we discuss methods to combine sensor data representing heterogeneous time averages in order to i) reconstruct an estimate and associated uncertainty of signal at any given time, and ii) use data from reference sensors to calibrate low cost sensors *in situ*.

1 Introduction

Many monitoring instruments gather time-averaged data. For example, the London Air Quality Network of reference sensors [6] provides 15 minute or hourly averages. Often we wish to combine knowledge gathered from instruments using different time-windows, e.g., reference sensors providing hourly averages along with low-cost sensor providing less accurate data at finer time resolutions. There are often problems with how to interpret time-averaged data where it is known that some of the averages are taken over a partial period due to missing data. We may also want a construct the signal in finer time-resolved steps in order to make comparisons with other variates. In this paper, we present a general approach for constructing a signal from time-averaged data. We make almost no assumptions about the regularity of the averaging process and the algorithm works for data representing averages over completely random sets of time steps. Instead, we make an assumption that at least part of the underlying process evolves smoothly so that the responses at nearby times are correlated to some extent. In section 2 we describe a general model for time-averaged data related to a temporally-correlated signal and in section 3 we describe how the model parameters can be determined from time-averaged data. In section 4, we illustrate how the signal reconstruction process behaves on problems that occur in practice. Our concluding remarks are given in section 5.

2 General model for time-averaged data

We assume that a signal x_q at time t_q can be modelled as

$$\boldsymbol{x} = C\boldsymbol{a} + \boldsymbol{e} + \boldsymbol{\delta}, \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_q, \ldots x_M)^{\mathrm{T}}$ is the signal at times $\boldsymbol{t} = (t_1, \ldots, t_q, \ldots t_M)^{\mathrm{T}}$, $\boldsymbol{a}$ are parameters that describe expected characteristics of the system such as drift and/or cyclical behaviour, C is the observation matrix associated with $\boldsymbol{x}$ and is usually specified by basis functions evaluated at $\boldsymbol{t}$, $\boldsymbol{e}$ are temporally correlated effects [1,7] whose correlation is modelled in terms of a correlation kernel

$$\mathrm{cov}(e, e') = k(t, t'),$$

e.g.

$$k(t, t'|\sigma, \tau) = \sigma_E^2 \exp\left\{-\frac{1}{2\tau^2}(t - t')^2\right\}, \tag{2}$$

and $\boldsymbol{\delta}$ are uncorrelated random effects $\boldsymbol{\delta} \in \mathrm{N}(\boldsymbol{0}, \sigma_R^2 I)$. We let $V_{\boldsymbol{e}}$ be the $M \times M$ variance matrix associated with $\boldsymbol{e}$ calculated using $\boldsymbol{t}$ and assume that

$$\boldsymbol{e} \in \mathrm{N}(\boldsymbol{0}, V_{\boldsymbol{e}}), \quad V_{\boldsymbol{e}} = V_{\boldsymbol{e}}(\boldsymbol{t}|\sigma_E^2, \tau). \tag{3}$$

We also assume that the times $\{t_1, t_2, \ldots, t)M\}$ are regularly spaced.

We assume that a number of time-averaged measurements $\boldsymbol{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$ are available. We let $\boldsymbol{t}_i \subset \{t_1, \ldots, t_M\}$ be the times associated with the ith measurement, n_i the number elements in $\boldsymbol{t}_i$, and $\boldsymbol{d}_i$ be the $M \times 1$ vector such that d_q/n_i if t_q is an element of $\boldsymbol{t}_i$ and is zero otherwise. Thus, the vector $\boldsymbol{d}_i$ can be used to construct the average $\boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{x}$ of the signal $\boldsymbol{x}$ over the times $\boldsymbol{t}_i$. The ith measurement is modelled as

$$y_i = \boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{x} + \epsilon_i, \quad \epsilon_i \in \mathrm{N}(0, \sigma_i^2). \tag{4}$$

We denote by $V(\boldsymbol{\sigma})$ the diagonal variance matrix with σ_i^2 in the ith diagonal element. If D is the $m \times M$ matrix with $\boldsymbol{d}_i^{\mathrm{T}}$ in the ith row, then (4) can be written in matrix-vector form

$$\boldsymbol{y} = D\boldsymbol{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \in \mathrm{N}(\boldsymbol{0}, V(\boldsymbol{\sigma})). \tag{5}$$

In most circumstances, m, the number of time-averaged data recorded, will be significantly smaller than M, the complete set of times. The goal of the analysis is to estimate the signal $\boldsymbol{x}$ based on the measurements

[*] e-mail: alistair.forbes@npl.co.uk

Article available at http://cfmetrologie.edpsciences.org or https://doi.org/10.1051/metrology/201712007

$\boldsymbol{y}$ gathered according to (4), and the prior correlation information about $\boldsymbol{e}$ specified by (3).

The formulation in (5) is quite general. The matrix D can represent any set of time averaging so that a quite heterogeneous set of sensors can be modelled in this way. For example, some sensors providing monthly averages do so in either four or five week time intervals, depending on the length of the month. There is no requirement that the vectors $\boldsymbol{t}_i$ (and therefore D) represent a sets of contiguous time steps, although in most practical applications they will be. The matrix $V(\boldsymbol{\sigma})$ allows for different sensor accuracies to be represented, e.g., an accurate (expensive) reference sensor giving hourly averages along side a (low cost) sensor giving ten minute averages with much less accuracy. The model can be expanded to incorporate parameters $\boldsymbol{b}$ that represent calibration parameters for the less accurate sensors [3] so that the reference sensor can be used to calibrate the less accurate sensors *in situ*. A simple example of this type of co-calibration is given below.

3 Estimation of the model parameters

Let D be the $m \times M$ matrix with $\boldsymbol{d}_i^{\mathrm{T}}$ in the ith row and $L_{\boldsymbol{e}}$ the Cholesky [4] of factor of $V_{\boldsymbol{e}} = L_{\boldsymbol{e}} L_{\boldsymbol{e}}^{\mathrm{T}}$ and introduce the parameters $\boldsymbol{f}$ related to $\boldsymbol{e}$ through $\boldsymbol{e} = L_{\boldsymbol{e}}\boldsymbol{f}$ so that (1) and (4) can be combined to form

$$\boldsymbol{y} = DL_{\boldsymbol{e}}\boldsymbol{f} + DC\boldsymbol{a} + D\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{6}$$

with

$$\boldsymbol{\delta} \in \mathrm{N}(\boldsymbol{0}, \sigma_R^2 I), \quad \boldsymbol{\epsilon} \in \mathrm{N}(\boldsymbol{0}, V(\boldsymbol{\sigma})), \quad \boldsymbol{f} \in \mathrm{N}(\boldsymbol{0}, I).$$

Re-defining $\boldsymbol{\epsilon}$ to be the sum of the random effects $D\boldsymbol{\delta} + \boldsymbol{\epsilon}$, we end up with

$$\boldsymbol{y} = DL_{\boldsymbol{e}}\boldsymbol{f} + DC\boldsymbol{a} + \boldsymbol{\epsilon}, \tag{7}$$

with

$$\boldsymbol{f} \in \mathrm{N}(\boldsymbol{0}, I), \quad \boldsymbol{\epsilon} \in \mathrm{N}(\boldsymbol{0}, V(\boldsymbol{\sigma}) + \sigma_R^2 DD^{\mathrm{T}}).$$

Let $L(\boldsymbol{\sigma}, \sigma_R)$ be the Cholesky factor of

$$V(\boldsymbol{\sigma}) + \sigma_R^2 DD^{\mathrm{T}} = L(\boldsymbol{\sigma}, \sigma_R) L^{\mathrm{T}}(\boldsymbol{\sigma}, \sigma_R).$$

and define the weighting matrix $W = W(\boldsymbol{\sigma}, \sigma_R)$ by

$$W(\boldsymbol{\sigma}, \sigma_R) = L^{-1}(\boldsymbol{\sigma}, \sigma_R).$$

Estimates $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{a}}$ of $\boldsymbol{f}$ and $\boldsymbol{a}$, respectively, are found by solving the least squares system

$$\check{C}\check{\boldsymbol{a}} \approx \check{\boldsymbol{y}}, \tag{8}$$

where

$$\check{C} = \begin{bmatrix} WDL_{\boldsymbol{e}} & WDC \\ I & \end{bmatrix}, \quad \check{\boldsymbol{a}} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{a} \end{bmatrix}, \quad \check{\boldsymbol{y}} = \begin{bmatrix} W\boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}.$$

The use of the weighting matrix W ensures that the variance matrix associated with the data $\check{\boldsymbol{y}}$ is the identity matrix. This means that the variance associated with the estimates $[\hat{\boldsymbol{a}}^{\mathrm{T}} \hat{\boldsymbol{f}}]^{\mathrm{T}}$ is given by

$$V_{\check{\boldsymbol{a}}} = \left(\check{C}^{\mathrm{T}}\check{C}\right)^{-1}.$$

The estimated signal is given by $\hat{\boldsymbol{x}} = C\hat{\boldsymbol{a}} + L_{\boldsymbol{e}}\hat{\boldsymbol{f}}$ with associated variance matrix

$$V_{\hat{\boldsymbol{x}}} = [L_{\boldsymbol{e}}\ C] V_{\check{\boldsymbol{a}}} [L_{\boldsymbol{e}}\ C]^{\mathrm{T}}.$$

4 Numerical examples

4.1 Reconstruction of a signal from hourly averages

We first illustrate the behaviour of the algorithm in addressing the problem of determining estimates of a signal at every five minutes given hourly averages. The model involves parameters $\boldsymbol{a}$ modelling a linear trend. The simulation data was generated as in equation (6), with the independent random component $\boldsymbol{\delta}$ generated with $\sigma_R = 0.01$ and the temporally correlated component $\boldsymbol{e}$ generated as in (3) using the kernel in (2) with $\sigma_E = 0.1$ and correlation times scale parameter $\tau = 1.00$, 0.50 and 0.25. The hourly averages are generated as in (4) with $\sigma_i = 0.005$. The unit for the times is 1 hour. The signal is in arbitrary units for this simulation. The three values of τ correspond to a correlation between the signal at one time and at one hour later (or earlier) being 0.61, 0.14 and 0.000 3, respectively.

Figure 1 shows the reconstructed signals for data corresponding to the three values of τ over a 12 hour period. For the case $\tau = 1.00$ (top graph), the reconstructed signal is a very good representation of the actual signal over the period. Considering that we are estimating the 144 values of the signal x_i (every five minutes over 12 hours, along with the two linear drift parameters $\boldsymbol{a}$, from just 12 measurements, the quality of the reconstruction is very good. The reason for this is that the temporal correlation is forcing a degree of smoothness on the signal and the hourly averages are sufficient to select a good reconstruction from all the possible signals with the requisite degree of smoothness.

A more quantitative argument can be derived by looking at the eigenvalues associated with the variance matrix $V_{\boldsymbol{e}}$. The sum of the eigenvalues is equal to the trace of $V_{\boldsymbol{e}}$, the sum of the diagonal elements of the variance matrix. For the case $\tau = 1.00$, the largest 10 eigenvalues account for approximately 98 % of the trace of $V_{\boldsymbol{e}}$, indicating that $V_{\boldsymbol{e}}$ can be approximated well by a rank 10 matrix. Or, in other words, there is a small number of effective degrees of freedom associated with the model [2,5] so that the 12 hourly averages determine good estimates of these parameters. The reconstructed signal explains over 98 % of the variance of the data.

The reconstruction for the case $\tau = 0.50$ is given in the middle graph of figure 1. For this case, it is clear that the reconstruction does not capture all elements of the signal. The largest 10 eigenvalues of $V_{\boldsymbol{e}}$ account for approximately 80 % of its trace while the reconstructed signal accounts for approximately 78 % of the variance of the data. The bottom graph in figure 1 is the reconstruction for the case $\tau = 0.25$. For this case, largest 10 eigenvalues of $V_{\boldsymbol{e}}$ account for approximately 48 % of its trace while the reconstructed signal accounts for approximately 47 % of the variance of the data. From another point of view, there is a much broader range of signals with the same smoothness characterisation that could have given rise to the measured hourly averages.

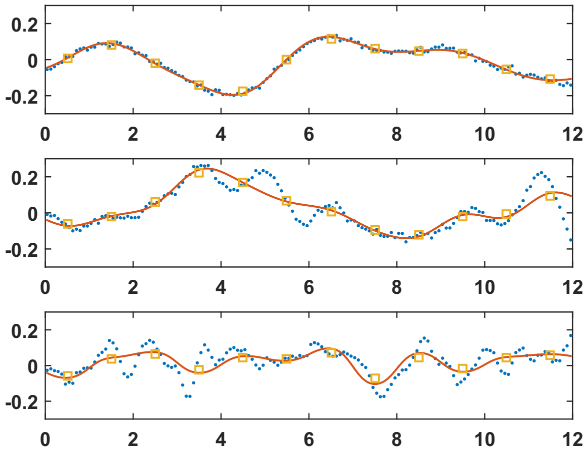

Fig. 1. Reconstruction (solid line) of a simulated signal sampled at 5 minute intervals (dots) from hourly averages (squares) for data generated with timescale parameter $\tau = 1.00$ (top), 0.50 (middle) and 0.25 (bottom). The unit associated with times is 1 hour. The signal is in arbitrary units.

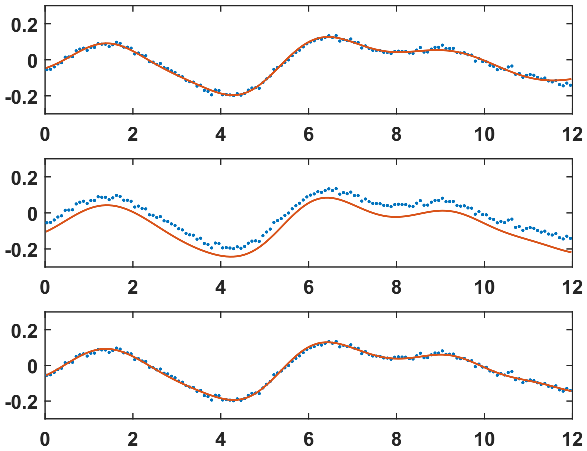

Fig. 2. Reconstruction (solid line) of a simulated signal sampled at 5 minute intervals (dots) from i) hourly averages from a reference sensor (upper) ii) 10 minute averages from a low-cost sensor with a calibration offset (middle), and ii) both sets of sensor data (lower), for data generated with timescale parameter $\tau = 1.00$. The unit associated with times is 1 hour. The signal is in arbitrary units.

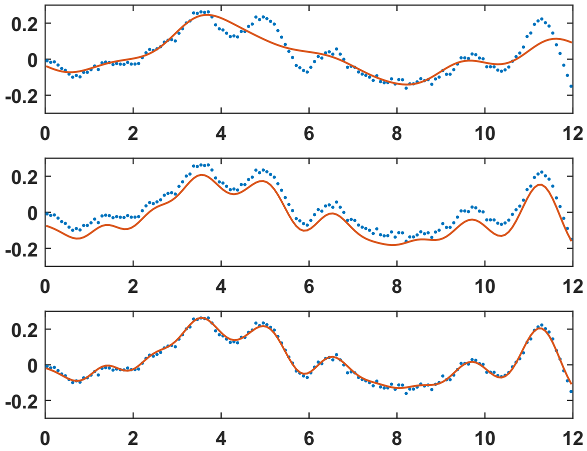

Fig. 3. As fig. 2 but for $\tau = 0.50$.

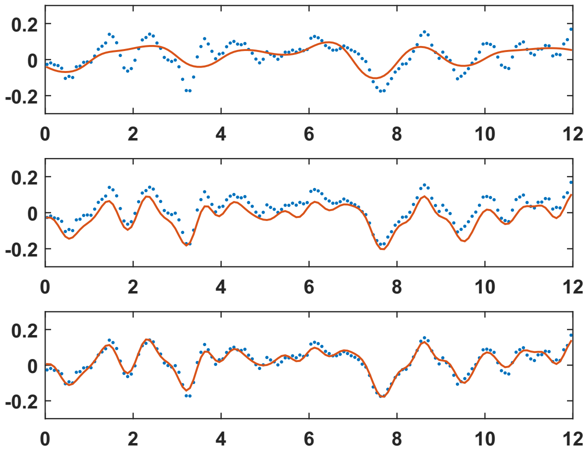

Fig. 4. As fig. 2 but for $\tau = 0.25$.

4.2 Collaborative measurement using a reference and low-cost sensor

The second set of simulations relates to the very practical requirement of how to combine less accurate, low-cost sensors with a reference sensor in order to deliver an enhanced capability. The simulations involve exactly the same set of signals and hourly averages as described above in section 4.1. However, 10 minute averages are also available from a low cost sensor that provide measurements y_i related to $\boldsymbol{x}$ according to

$$y_i = b + \boldsymbol{d}_i^{\mathrm{T}} \boldsymbol{x} + \epsilon_i, \quad \epsilon_i \in \mathrm{N}(0, \sigma_L^2), \tag{9}$$

where b is an offset common to all the measurements and ϵ_i is a random effect with associated uncertainty σ_L. The low-cost nature of the sensor is modelled in this case by the offset, with modest prior information $b \sim \mathrm{N}(0, \sigma_O^2)$, with $\sigma_O = 0.1$, and the fact that $\sigma_L = 0.025$ is five times greater than $\sigma_i = 0.005$ for the reference sensor. This additional sensor can be accounted for in the general model (6) through appropriate assignment of the matrices C, D and $V(\boldsymbol{\sigma})$.

Figure 2 shows in the upper graph the reconstructed signal of a simulated signal derived from the reference

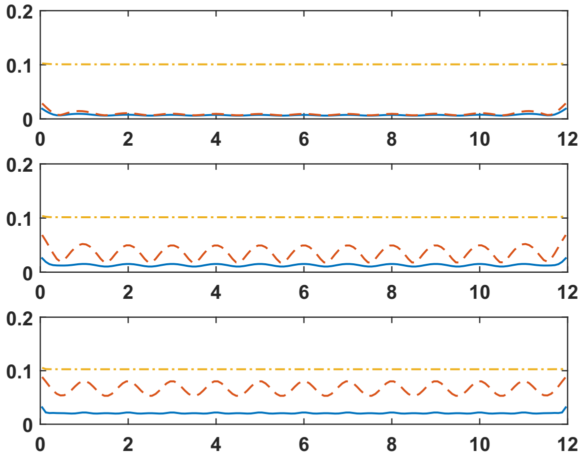

Fig. 5. Standard uncertainties associated with the reconstructed signals determined from i) reference sensor data alone (dashes), ii) low-cost sensor data alone (dot-dashes), and iii) both sets of data (solid) for the cases $\tau = 1.00$ (upper graph), $\tau = 0.5$ (middle) and $\tau = 0.25$ (lower).

data alone, for the case $\tau = 1.00$. This signal is exactly the same as that in the upper graph of figure 1. The middle graph in figure 2 is the signal reconstructed from the low-cost sensor data alone. The reconstruction is seen to follow the data well but is offset due to the systematic effect b. The uncertainty associated with this systematic effect determined from the low-cost data alone is exactly the same as the prior uncertainty σ_O: the data provides no extra information about b. The lower graph in figure 2 is the signal reconstructed from both the reference and low-cost sensor data and gives a good reconstruction of the data, as in the upper graph. The standard uncertainty associated with b derived from the combined data set is $u(b) = 0.004$, comparable with the standard deviation $\sigma_i = 0.005$ of the random effects associated with the reference sensor, showing that the analysis of the combined set of data gives an accurate *in situ* calibration of the low-cost sensor.

The upper graph in figure 5 gives the standard uncertainty associated with the three reconstructed signals. The uncertainty associated with signal reconstructed from the low-cost sensor alone is dominated by the uncertainty of 0.1 associated with the offset b. The uncertainties associated with the other reconstructions are similar with the combined data set giving marginally lower uncertainties, as expected. In a practical setting, the low-cost sensor is providing little useful information if the background signal is sufficiently smooth (but the collaborative approach provides the *in situ* calibration).

Figure 3 provides similar reconstructions as figure 2 but for the case $\tau = 0.5$, representing less smooth signals. Here, the main outcome is that the collaborative approach, in the lower graph, shows a good reconstruction of the signal, significantly better than that derived from the reference (upper) or low-cost sensor data (middle) alone. This fact is also reflected in the standard uncertainties associated with the reconstructed signals with the uncertainties associated with the collaborative approached significantly smaller that those for the other two reconstructed signals.

Figure 4 shows the reconstructions for the case $\tau = 0.25$ and again shows the collaborative approach gives a good reconstruction (lower graph). The benefit of the collaborative approach, in terms of the standard uncertainties in the reconstructed signals, is seen in the lower graph of figure 5.

For all three values of τ, the standard uncertainty $u(b)$ associated with the estimate of b derived from the combined data sets is $u(b) = 0.004$, demonstrating that the collaborative approach can determine an accurate *in situ* calibration of the low-cost sensor.

5 Concluding remarks

We have described a general algorithm for reconstructing a signal from time-averaged data and illustrated its performance on simulated data representing practical problems in air quality monitoring, for example. The underlying model uses a prior assumption of temporal correlation in the signal in order to improve the reconstruction. The algorithm can be used to estimate the signal at finer time resolutions, for combining multiple data streams that are averaged over different time windows, and providing the *in situ* calibration of sensors against a reference sensor.

Acknowledgement

This work was funded by the UK's National Measurement System Programme. I thank my colleague Kavya Jagan for comments on an earlier draft of this paper.

References

1. N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data.* Wiley, Hoboken, New Jersey, 2011.
2. A. B. Forbes. Empirical functions with pre-assigned correlation behaviour. In F. Pavese, W. Bremser, A. Chunovkina, N. Fischer, and A. B. Forbes, editors, *Advanced Mathematical and Computational Tools for Metrology X*, pages 17–28, Singapore, 2015. World Scientific.
3. A. B. Forbes. Traceable measurements using sensor networks. *Transactions on Machine Learning and Data Mining*, 8(2):77–100, October 2015.
4. G. H. Golub and C. F. Van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, third edition, 1996.
5. T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning.* Springer, New York, 2nd edition, 2011.
6. L. Mittal and G. Fuller. London Air Quality Network: Summary Report 2016. Technical report, King's College, London, June 2017.
7. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, Cambridge, Mass., 2006.