ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Blind source separation problem in GPS time series

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Gualandi, A., Serpelloni, E., Belardinelli, M. (2016). Blind source separation problem in GPS time series. JOURNAL OF GEODESY, 90(4), 323-341 [10.1007/s00190-015-0875-4].

Availability:
This version is available at: https://hdl.handle.net/11585/538514 since: 2021-12-14

Published:
DOI: http://doi.org/10.1007/s00190-015-0875-4

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

12 October 2024

This is the final peer-reviewed accepted manuscript of:

Gualandi, A., Serpelloni, E. & Belardinelli, M.E. Blind source separation problem in GPS time series. Journal of Geodesy 90, 323–341 (2016).

The final published version is available at: https://doi.org/10.1007/s00190-015-0875-4

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Manuscript Reviewed Click here to download Manuscript Manuscript_reviewd.docx

Click here to view linked References

Blind Source Separation problem in GPS time series

A. Gualandi[1,3,], E. Serpelloni[2] and M.E. Belardinelli[3]*

1 Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna

2 Istituto Nazionale di Geofisica e Vulcanologia, Centro Nazionale Terremoti

3 Dipartimento di Fisica e Astronomia, Settore di Geofisica, Università di Bologna

* Corresponding author: Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Via D. Creti, 12, 40128 Bologna (IT). Phone: +39 051 4151474, e-mail: *adriano.gualandi@ingv.it*

Abstract

A critical point in the analysis of ground displacement time series, as those recorded by space geodetic techniques, is the development of data driven methods that allow the different sources of deformation to be discerned and characterized in the space and time domains. Multivariate statistic includes several approaches that can be considered as a part of data-driven methods. A widely used technique is the Principal Component Analysis (PCA), which allows us to reduce the dimensionality of the data space while maintaining most of the variance of the dataset explained. However, PCA does not perform well in finding the solution to the so-called Blind Source Separation (BSS) problem, i.e. in recovering and separating the original sources that generates the observed data. This is mainly due to the fact that PCA minimizes the misfit calculated using a L_2 norm (χ^2), looking for a new Euclidean space where the projected data are uncorrelated. The Independent Component Analysis (ICA) is a popular technique adopted to approach the BSS problem. However, the independence condition is not easy to impose, and it is often necessary to introduce some approximations. To work around this problem, we test the use of a modified variational bayesian ICA (vbICA) method to recover the multiple sources of ground deformation even in the presence of missing data. The vbICA method models the probability density function (pdf) of each source signal using a mix of Gaussian distributions, allowing for more flexibility in the description of the pdf of the sources with respect to standard ICA, and giving a more reliable estimate of them. Here we present its application to synthetic GPS position time series, generated by simulating deformation near an active fault, including inter-seismic, co-seismic, and post-seismic signals, plus seasonal signals and noise, and an additional time dependent volcanic source. We evaluate the ability of the PCA and ICA decomposition techniques in explaining the data and in recovering the original (known) sources. Using the same number of components, we find that the vbICA method fits the data almost as well as a PCA method, since the χ^2 increase is less than 10% the value calculated using a PCA decomposition. Unlike PCA, the vbICA algorithm is found to correctly separates the sources if the correlation of the dataset is low (<0.67) and the geodetic network is sufficiently dense (10 cGPS stations within a box

of side equal to two times the locking depth of a fault where an earthquake of $M_w > 6$ occurred). We also provide a cookbook for the use of the vbICA algorithm in analyses of position time series for tectonic and non tectonic applications.

Keywords: Time series analysis, Transient deformation, Seismic cycle, Global Positioning System (GPS), Variational Bayesian Independent Component Analysis (ICA, vbICA), Principal Component Analysis (PCA).

1 - Introduction

Often Earth scientists try to increase the knowledge about a specific process just increasing the amount of data collected. Nevertheless, the extraction of the most significant information from a huge and messy data set can be not trivial. Most of time, huge data sets contain the evolution of a certain observable w.r.t. time, arranged in the form of time series. This is true for geodetic measurements of crustal deformation. In the last decade the number of space geodetic data available for studying Earth's surface dynamics grew remarkably, due to the increase of Global Navigation Satellite System (GNSS) networks and the availability of new Synthetic Aperture Radar (SAR) observations. In particular, three fundamental aspects improved: the spatial coverage, the temporal coverage, and the accuracy of the measurements. The ability to detect ground displacements with accuracy down to the millimeter level enables a better understanding of deformations associated to volcanic processes (e.g., Owen et al., 2000), as well as a better characterization of the different processes occurring at active fault zones, including a better estimate of the inter-seismic linear velocities, tectonic strain loading, and fault movements (slips) during the different phases of the earthquake cycle. Examples of tectonic deformation events are co-seismic and post-seismic slip (e.g., Johanson et al., 2006, Perfettini et al., 2010), Slow Slip Events (SSE), i.e. slip events not detected by seismometers and having a much longer duration than ordinary earthquakes of comparable seismic moment, and pre-seismic slip events (e.g., Heki et al., 1997; Ito et al., 2013).

Post-seismic slip, SSE, and pre-seismic slip events are all examples of transient deformations. In general, a transient deformation signal is "a nonperiodic, nonsecular accumulation of strain in the crust" (Riel et al., 2014), and its origin can be tectonic (e.g., post-seismic and SSE deformations) as well as non-tectonic (e.g., tidal loading, hydrological loading, anthropogenic processes). The detection and characterization of transient events on faults is a fundamental task of tectonic geodesy, with important implications for the evaluation of the seismic hazard, and several approaches have been developed in recent times. Lohman and Murray (2013), for example, described the results of some of these different methods, ranging from the visual inspection to more refined image processing techniques, Principal Component Analysis (PCA), Kalman filtering with spatial basis functions, space-time correlations, and so on, which have been applied in the framework of the SCEC blind transient detection project (http://collaborate.scec.org/transient, last access June 9, 2015).

We can classify the methods developed for the analysis of ground displacement time series data in two main categories. The first one consists in determining the parameters of a pre-determined model assumed to explain the data, where the pre-determined model is made up of analytic functions expressing a presumed contribution to the observed temporal evolution. A linear combination of these different functions results in the final model of the time series. In general, the most commonly used model considers at least the following contributions: a linear (secular) trend, cyclic and seasonal signals, and offsets (including instrumental and co-seismic offsets). It is also possible to add some other functional forms in order to describe additional transient signals (e.g., exponential or logarithmic functions in order to describe the post-seismic decay). A clear advantage of this approach is that each contribution is associated to a given (known) physical process. On the other hand, we are mostly interested in understanding what is not already known, i.e. those signals for which we do not have any well established pre-determined model. We can classify as *model-based* those techniques that rely on the estimate of the best parameters of a given dictionary of functions. Riel et al. (2014), for example, have proposed to use a dictionary of non-orthogonal functions to mimic the ground displacements, and implemented an automatic routine in order to detect the

dominant timescales and onset times of transient signals, i.e. signals that are not explained with the basic model described above (i.e., linear + cyclic + offsets). The availability of a wide dictionary allowed them to reconstruct transient signals of different shapes. Considering the time series one-by-one, the temporal correlated colored noise causes several false detection throughout the time series. In order to work around this problem they introduced a spatial sparsity weighting approach.

The second method for the analysis of ground displacement time series, which also provides an alternative way to mitigate the temporal correlated noise issue, consists in *multivariate statistical* techniques, which consistently exploits the information contained in data recorded by a network of sensors. This second main category of time series analysis encompasses a wide spectrum of different methods, among which the most popular one is the PCA. This technique has been used for the dimensionality reduction problem in order to extract the most relevant part of the data (relevance depends on the specific application). Such a reduction consists in projecting the data onto a new coordinate system, given by the Principal Components (PCs). The new reference system is linear, and the new axes (i.e., the PCs) are orthogonal. This projection allows an easier interpretation of even a huge amount of data in terms of few components. The PCA technique has been widely applied in geodesy, both to filter common mode noise in GPS networks (e.g., Dong et al., 2006) and to detect regional tectonic signals (e.g., Kositsky and Avouac, 2010; Ji and Herring, 2011). The PCA uses only statistics up to the second moment, i.e. the variance, diagonalizing the covariance matrix of the data in order to decorrelate the original dataset. This implies that the assumption underlying PCA is that the data projected onto the components are normally distributed. Let us suppose that the data are the result of a mix of different ongoing processes. In signal processing theory each process is referred to as a *source signal*, and the observed time series are the *sensor signals*. This means that the sensor outputs consist of a mix of the source signals. Figure 1 shows the pdf distributions for some of the sources most commonly present in GPS time series, such as a linear trend, an annual signal, and a post-seismic signal. Clearly, the pdf for these sources are not Gaussian, preventing a proper representation via a single PC. Then, in the case of a mix of such sources, the PCs represent

only a combination of the true physical sources beneath the observed data. In other words, in this case the PCs do not have any physical meaning if individually taken.

To make possible a physical interpretation of the components, it becomes of fundamental importance to find out what the original source signals are, making the fewest (and the most reasonable) number of assumptions as possible. This is the goal of the Blind Source Separation (BSS) problem. One of the most popular approaches to solve the BSS is the Independent Component Analysis (ICA). This multivariate technique still remains in the field of linear decompositions. However, the data are projected onto a system of coordinates where each component is no longer constrained to be orthogonal to another one. In other words, the Independent Components (ICs) constitute a non-orthogonal basis for this reference system. Only few applications of ICA techniques to geodetic data have been presented so far in the literature. For example, Bottiglieri et al. (2007) have applied the FastICA algorithm (Hyvärinen and Oja 1997) to a cGPS network located in the Neapoletan volcanic area. Forootan and Kusche (2012, 2013) have applied a modification of the JADE algorithm (Cardoso and Soulomiac, 1993) to the Gravity Recovery and Climate Experiment (GRACE) data. They have shown that rotating the experimental orthogonal functions perfectly separates an unknown mixture of trend and sinusoidal signals present in the data, provided that the length of the dataset is infinite. However, as previously said, we are mostly interested in detecting and characterizing also transient signals in geodetic time series (e.g., Lohman and Murray, 2013 and references therein). If pdfs of transient signals are non-unimodal (as in Figure 1), then classical ICA algorithms do not always perform an optimal decomposition (e.g., Choudrey, 2002, Section 2.4 and references therein). In those cases Choudrey (2002) has shown that the variational bayesian ICA (vbICA) is a more flexible method to solve the BSS problem.

In this work we apply the vbICA approach to synthetic position time series. In particular, the original approach is suitably modified in order to deal with missing data in time series. We simulate position time series from a GPS network recording in proximity of an active fault and a volcanic source. All the tests are performed using the so-called static approach, i.e. assuming that the mix of

different ongoing processes (or source signals) is independent of time. In other words the mix is supposed to be instantaneous and the sources are supposed to be fixed in space, i.e. they are not moving from their starting position. In general, this could be not the case, but it is still a good approximation for the geophysical scenarios investigated here. Owing to its mathematical complexity, the moving sources BSS problem is still under investigation, and facing it is beyond the goals of the present work. In the following we present a summary of the theory behind the vbICA approach (Sections 2, 3, and 4), we describe the way we generated synthetic data (Section 5), and the results of the analysis of synthetic data with vbICA and other more popular multivariate statistical techniques (Section 6).

2 – Principle of ICA

Let us consider a network of N continuous GPS stations, for which we have daily positions at each station. Since GPS measurements are three-dimensional, for each station we have 3 time series: for the east, north, and vertical components. If $M = 3N$ is the total number of time series, and T is the total number of recorded epochs, the data matrix corresponding to the position of the GPS stations is:

$$\mathbf{X}_{M\times T} = \begin{pmatrix} x_{11} & \cdots & x_{1T} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MT} \end{pmatrix} \tag{1}$$

where x_{jt} is the position of the time series j at epoch t, with $j = 1, \ldots, M$ and $t = 1, \ldots, T$. In the framework of multivariate statistical techniques the set of data collected by a sensor at different times corresponds to a sample that describes a random variable (rv) associated to the specific time series. This means that the matrix $\mathbf{X}$ corresponds to a vector of M rvs, described by the samples of the M time series.

Suppose that the data are generated by few sources ($L < M$) and that linearly combining these few sources we can reconstruct the observed data, and assume that the processes related to the different sources can be studied separately. This means that we assume that the processes are mutually

independent. Finally, let us suppose that some Gaussian noise is perturbing the measurements. These assumptions can be summarized as follows:

$$\mathbf{X}_{M\times T} = \mathbf{A}_{M\times L}\mathbf{S}_{L\times T} + \mathbf{N}_{M\times T} \tag{2}$$

where **A** is called mixing matrix, **S** source matrix, and **N** noise matrix. Each row of **S** contains the temporal evolution associated to a given source, and the sources are statistically independent one from the other. This corresponds to describe the M observed rvs using a linear combination of only L variables, whose pdfs describe the temporal evolution associated to each row of **S**. For the independence among the sources the joint pdf of the L rvs is necessarily factorized as:

$$p(s_1, \dots, s_L) = \prod_{i=1}^{L} p(s_i) \tag{3}$$

where $p(s_i)$ are the pdfs of each source s_i, with $i = 1, \dots, L$. In order to identify **S** under this constraint, two possible strategies can be pursued. The first methodology is called *mapping* approach: it looks for a function $\Upsilon: \mathbb{R}^M \rightarrow \mathbb{R}^L$ that performs a projection from the M-dimensional data space to the L-dimensional source space. A second methodology is the *modeling* approach: it looks for a function $\Upsilon: \mathbb{R}^L \rightarrow \mathbb{R}^M$ that allows the data to be reproduced using a generative model. The mapping approach is the first that has been developed, and it is the most commonly used (e.g., Comon, 1994; Hyvärinen and Oja, 1997; Cardoso and Souloumiac, 1993). Different algorithms have been proposed, which basically build a contrast function to be minimized. Such a contrast function expresses how far from being independent are the sources, i.e. how far the right hand side of equation (3) is from the left hand side. Since the actual sources are unknown, some approximations are needed. The approximations usually involve 4th order cumulants and are based on the definition of mutual information or negentropy (e.g., Comon, 1994; FastICA by Hyvärinen and Oja, 1997). The modeling approach instead creates a generative model, and looks for the model parameters that allow the data to be better explained. The contrast function to be maximized is the likelihood or, in a bayesian

framework, the posterior pdf of the parameters. Roberts and Everson (2001) have shown that a generative model for all the possible mapping procedures exists. Here we use the generative model approach, and we use the notation of Choudrey (2002).

In order to take into account missing data, which is a common problem in GPS time series, we have modified the variational bayesian ICA (vbICA) code of Choudrey (2002) (http://www.robots.ox.ac.uk/~parg/projects/ica/riz/code.html, last access June 10, 2015) following Chan et al. (2003). The modifications consist in applying a mask of 0 (missing) and 1 (recorded) to the data and to the formulas used to update the parameters of the generative model. Here we present a short overview of the main concepts, and for more details see Section S1 of the Supplementary material.

3 – Description of vbICA

A generative model $\mathfrak{M}$ is characterized by some observed variables (**X**), some hidden or latent variables (**H**), some hidden parameters (**Θ**), and the mutual relationships between all these quantities. The observations and the hidden variables are quantities identified with the "real world". In order to have the model working, some parameters exist and their distributions are modeled using further parameters (called hyper-parameters). Both the parameters and the hidden variables are unknown, and they are indicated as "weights", **W** = {**H**, **Θ**}. The prior pdf of the weights, given a model $\mathfrak{M}$, is indicated as $p(\mathbf{W}|\mathfrak{M})$. The goal of a generative model is to find the best weights in order to explain the observations and match the a priori knowledge, embodied in the particular structure of the model $\mathfrak{M}$ and the hyper-parameter values of the prior pdfs. In a bayesian framework, given a model $\mathfrak{M}$ and the observed data **X**, maximizing the posterior pdf over weights **W** given the data **X** is the best choice for **W**:

$$p(\mathbf{W}|\mathbf{X}, \mathfrak{M}) = \frac{p(\mathbf{X}|\mathbf{W},\mathfrak{M})p(\mathbf{W}|\mathfrak{M})}{p(\mathbf{X}|\mathfrak{M})} \tag{4}$$

where the denominator is called the *evidence* for $\mathfrak{M}$ and is expressed by:

$$p(\mathbf{X}|\mathcal{M}) = \int p(\mathbf{X}|\mathbf{W},\mathcal{M})p(\mathbf{W}|\mathcal{M})\mathrm{d}\mathbf{W} \tag{5}$$

In practice, the computation of the integral in equation 5 is intractable in most of the cases, since it has to be calculated in the whole weight space, and also in this case approximations are needed in order to evaluate it. Choudrey (2002) proposed a variational approximation, which allows us to use the Negative Free Energy (NFE) as a contrast function to be maximized in order to find an approximating posterior pdf for the weights ($p'(\mathbf{W}|\mathcal{M})$) such that the divergence between the true ($p(\mathbf{W}|\mathbf{X},\mathcal{M})$) and the approximate posteriors is minimized. As shown in the Supplementary Material, the NFE can be expressed as:

$$NFE[\mathbf{X}] = \langle \ln(p(\mathbf{X},\mathbf{W}))\rangle_{p'(\mathbf{W})} + H[\mathbf{W}] \tag{6}$$

where the dependence on $\mathcal{M}$ is dropped for conciseness, $<\cdot>_{p'(\mathbf{W})}$ is the expected value given the pdf $p'(\mathbf{W})$, and $H[\mathbf{W}]$ is the entropy of $p'(\mathbf{W})$. In order to apply the variational approximation it is thus necessary to choose a particular form for the approximating pdf of the weights, $p'(\mathbf{W})$. The most common restriction on $p'(\mathbf{W})$ is that it factorizes into $\prod_{i=1}^{N} p'(w_i)$ for some partition $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ of $\mathbf{W}$ (e.g., Ormerod and Wand, 2010). Each weight $\mathbf{w}_i$ is a rv that can be described by a given distribution, governed by some hyper-parameters, or by other rvs.

For the particular case of the BSS problem, we want to find those parameters (i.e., those weights $\mathbf{W}$) that can explain the data $\mathbf{X}$ under the framework of linear combination of independent source signals. The vbICA approach makes use of the following partition: $\mathbf{W} = \{\mathbf{A}, \mathbf{S}, \mathbf{\Lambda}, \mathbf{q}, \boldsymbol{\theta}\}$, where $\mathbf{A}$ are the rvs describing the mixing matrix, $\mathbf{\Lambda}$ are the rvs describing the precision (i.e., the inverse of the variance) associated to the noise, $\mathbf{S}$ are the rvs describing the sources. Each of these rvs can be described by a given distribution or by other rvs. In particular, each source s_i is expressed via a mix of m_i Gaussian distributions, with $i = 1, \ldots, L$. Using $q_i = 1, 2, \ldots, m_i$ as an indicator variable expressing which Gaussian component of the i-th source is chosen for generating s_i, the complete collection of all possible choices for q_i, $i = 1, 2, \ldots, L$ (source states), is denoted as $\mathbf{q} =$

$\{\mathbf{q}_1, \dots, \mathbf{q}_\mathbf{m}\}$, with $\mathbf{m} = \prod_{i=1}^{L} m_i$. The mix for the i-th source is described by the rvs $\boldsymbol{\theta}_i = \{\boldsymbol{\pi}_i, \boldsymbol{\mu}_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{\pi}_i$ is a vector with m_i components containing the probabilities that the q_i-th Gaussian with mean μ_{i,q_i} and precision β_{i,q_i} is chosen for explaining the source s_i.

The independence enforced by the factorization $\prod_{i=1}^{N} p'(w_i)$ allows us to use an Expectation-Maximization algorithm in order to solve for the NFE maximization problem. At the same time the constraint (3) is automatically fulfilled, since the M observed time series are considered generated by mixing L independent source signals. It is possible to rigorously derive the updating equations (also called learning rules) for all the involved weights. These equations are extensively described in Chapter 5, and Appendixes B and C of Choudrey (2002). In section S1 of the Supplementary material we report the learning rules with the modifications we have introduced in order to deal with missing data, following Chan et al. (2003) (see equations from S17 to S39).

The choice of the priors plays a major role in the determination of the final result, and this step is the only one where some a priori choice is required by users in running the vbICA method in time series analysis. In all the case studies we use starting values of the hyper-parameters that are rather loose, in order to let as much as possible the data to reveal their intrinsic structure. For more details on the selection of the pdfs governing the rvs involved and the prior parameters, see Section S1 of the Supplementary material.

4 – Operational differences between PCA and vbICA

Since PCA is a widely used technique in the analysis of geodetic time series, here we aim at performing some comparison between the PCA and vbICA algorithms in extracting signal of interest for geophysical studies. In this work we use the PCA technique incorporated in the PCA-based Inversion Method software (PCAIM, Kositsky and Avouac, 2010), which exploits the Srebro and Jaakkola (2003) decomposition algorithm, allowing the user to take into account missing data while performing the decomposition of the data matrix. In order to compare the PCA and ICA algorithms, it is necessary to assume a common normalization for the PCA eigenvectors and the ICA sources. It is also necessary to impose some constraints in order to have a unique PCA and ICA (e.g., Comon,

1994). A common approach used to implement the PCA decomposition consists in the Singular Value Decomposition (SVD), where the data matrix **X** is decomposed in the matrices **U**, **Σ**, and **V**, as **X** = **U Σ V**^{T}, where **U** and **V** are unitary matrices, while **Σ** is diagonal. Since it is always possible to decompose a matrix in a unit column norm matrix and a diagonal matrix, in order to compare the ICA to the SVD decomposition we rewrite the ICA formulation as follows:

$$\mathbf{X}_{\mathrm{ICA}} = \mathbf{AS} + \mathbf{N} = \mathbf{U}_{\mathrm{ICA}}\boldsymbol{\Sigma}_{\mathbf{A}_{\mathrm{ICA}}}\boldsymbol{\Sigma}_{\mathbf{S}_{\mathrm{ICA}}}\mathbf{V}_{\mathrm{ICA}}^{T} + \mathbf{N} = \mathbf{U}_{\mathrm{ICA}}\boldsymbol{\Sigma}_{\mathrm{ICA}}\mathbf{V}_{\mathrm{ICA}}^{T} + \mathbf{N} \tag{7}$$

where the columns of $\mathbf{U}_{\mathrm{ICA}}$ and $\mathbf{V}_{\mathrm{ICA}}$ are unit norm columns, but they are not orthogonal, and $\boldsymbol{\Sigma}_{\mathrm{ICA}}$ is a diagonal matrix. It is worth noting that, differently from PCA, ICA does not organize into a diagonal matrix the variances of the dataset, and thus we can not use a criterion based on a threshold of explained variance in order to select the most appropriate number of components to retain, as usually done for a PCA. To this aim, the bayesian framework allows us to use the Automatic Relevance Determination (ARD) method (e.g., MacKay, 1994). A strong confidence in the starting model (strong priors) implies a similarity between the posteriors and the priors. Instead, under weak priors the data play the most important role in guiding the learning of the posterior parameters, and an increasing number of components is used to fit (or over-fit) the data. The ARD method exploits the fact that each of the L columns of the mixing matrix is associated with one of the L sources. Instead of assigning a different precision to each element of the mixing matrix, let us associate only one precision value α_i , with $i = 1, \ldots, L$, to each column. Thus the mixing matrix depends on the set of parameters $\alpha = \{\alpha_1, \ldots, \alpha_L\}$, which defines how strong is the assumption that the mean value of the columns of the mixing matrix is zero. In other words, a certain α_i defines how relevant is the source i for the explanation of the data. A large value corresponds to a posterior over mixing matrix column i dominated by the prior density, effectively setting the elements of column i to zero. This will result in heavy suppression of the i-th source signal for the data explanation. By monitoring the variance of each source signal - or, equivalently, the values of α_i - the most likely number of sources supported by the observation data can be determined (Choudrey, 2002). In this work, in order to determine

the number of components to retain, we directly compare the variances associated to the posterior of each column i, and if the maximum variance is more than 10 times bigger than the smallest one, we consider the source associated with the latter as noise, and discard the last component.

5 – Synthetic data generation

In order to validate the algorithm described in Section 3, we perform tests on synthetic time series, realized in order to simulate continuous GPS (cGPS) observations, with the goal of evaluating the ability of the algorithm to retrieve sources that are known a priori. In order to generate the cGPS time series datasets, we create some source signals (**S**), then we perform a specific linear combination of the sources (i.e., we use a predetermined mixing matrix **A**), and finally we add some noise (**N**).

We aim to simulate daily ground displacement data around an active fault or near a volcanic region. The time spanned for the analysis is 5 years, and the number of cGPS stations is fixed to 20 with variable geometries of the network with respect to the fault, or volcanic, source. The data matrix is $\mathbf{X}_{M\times T} = \mathbf{X}_{60\times 1827}$ (considering also leap years). As a reference epoch we assume the starting time $T_{start} = 0$. All the data of the epochs that follow T_{start} are expressed in mm and estimate the displacement along east, north, and vertical directions relative to the first epoch. The simulation of the seismic cycle is performed considering only one planar fault as the main source of the tectonic deformation, which is modeled using a thrust fault simplified by a planar dislocation in an elastic and homogeneous half space (0.25 Poisson modulus). A rigidity modulus $\mu = 30$ GPa is used to estimate the equivalent magnitude of the final distribution of slip on the fault plane. By taking the fault geometry as fixed, we vary the area of the fault undergoing slip during different phases of the earthquake cycle, as well as the value of slip, with the goal of simulating different kinematic settings with different deformation rates at play. The volcanic activity is simulated using an inflating and deflating Mogi source, i.e. the deformation at the surface is computed from the formula of Anderson

(1936) and Mogi (1958), which consider a hydrostatic pressure inside a sphere having assigned radius and center depth, and embedded in a homogeneous, semi-infinite elastic body.

The tests are performed modifying two different conditions:

1) source intensities (signal-to-noise ratio, SNR, and signal-to-signal ratio, SSR, the latter being defined, in analogy with the SNR, as the ratio between the power of one signal and the power of a second signal; for more details, see section S2 of the Supplementary material).

2) network quality (sensor location w.r.t. the fault plane, and percentage of missing data);

It is worth noting that point 1) depends only on the sources, while point 2) relies only on the sensors.

5.1 – Description of the source signals

By using a linear function, a Heaviside function, and a logarithmic function as sources (detailed in Section S2 of the Supplementary Material) we mean to simulate the 1) inter-seismic, 2) co-seismic, and 3) post-seismic stages of the seismic cycle. Commonly, the (secular) linear trend observed in GPS time series is explained as due to long term, inter-seismic, relative motion of crustal blocks. The differences in trends (i.e., velocities) among sites, e.g. across a fault, are due to inter-seismic deformation and long-term slip in the deep portion of the fault plane. In this simplified model of the inter-seismic stage, the fault slips aseismically beneath a certain depth (e.g., Segall, 2010), defined as locking depth. In its shallow, brittle portion the fault is locked during the inter-seismic stage. Since we are simulating daily data it is not possible to see the evolution of the co-seismic rupture, and what is recorded is just a jump from the position before and after the earthquake, here modeled as an Heaviside step function. The use of a logarithmic function to represent the post-seismic process is less obvious than the model of co-seismic signals. The post-seismic deformation, in fact, can be driven by different processes, each of which following a characteristic evolution with time (e.g. Barbot and Fialko, 2010, and references therein). Here we decide to model afterslip following Marone at al. (1991), i.e. adopting a logarithmic function, and in particular assuming a constant decay time of 1 day. It is worth noting that this choice is not critical for the goal

of this work, since the algorithm does not take advantage of any particular choice of the model of the transient signal since it does not impose any further constraint on the shape of the recovered source function than its a priori knowledge.

We test three scenarios relative to the intensity of the tectonic signals, which differ in the amount and distribution of slip on the fault plane during the co- and post-seismic stages:

1) M_w^{co} = 6.85, # patches along strike = 15, # patches along dip = 9, $slip^{co}$ = 400 mm;

   M_w^{ps} = 6.57, # patches along strike = 15, # patches along dip = 7, $slip^{ps}$ = 200 mm;

2) M_w^{co} = 6.29, # patches along strike = 5, # patches along dip = 4, $slip^{co}$ = 400 mm;

   M_w^{ps} = 6.09, # patches along strike = 5, # patches along dip = 3+5, $slip^{ps}$ = 100 mm;

3) M_w^{co} = 5.94, # patches along strike = 3, # patches along dip = 4, $slip^{co}$ = 200 mm;

   M_w^{ps} = 5.80, # patches along strike = 3, # patches along dip = 3+5, $slip^{ps}$ = 60 mm;

where $slip^{co}$ ($slip^{ps}$) is the constant value of co-seismic (post-seismic) slip on mentioned patches, which are located along the fault plane as shown in Figure 2. The size of each patch is ~3.1 km along strike and ~3.7 km along depth. In the scenarios number 2) and 3) the post-seismic slip occurs both below and above the region of co-seismic slip. In all the configurations some regions of co-seismic and post-seismic slip overlap (see Figure 2). For the deep creeping section we use three different inter-seismic velocity scenarios: 2, 12, and 60 mm/yr, in order to span typical conditions of slow, intermediate, and fast deformation rates. We calculate the surface displacement resulting from fault slip using the Green's function computed from the solutions of Okada (1985).

We add also a seasonal signal characterized by a sinusoidal temporal evolution with a 1 yr period and a spatial Gaussian distribution, centered in the SW corner of the region considered (Figure 4). This can simulate, for example, the effect of a water reservoir located near the SW corner of region, where the signal amplitude is maximum (e.g., Argus et al., 2014).

Finally, in order to simulate a volcano-tectonic context, or the occurrence of a transient deformation superimposed on a post-seismic deformation, we use a Mogi source located at 5 km

depth, NW w.r.t. the fault (40 km West and 60 km North the surface projection of the fault top center). We use a linear combination of arctangent functions to reproduce $V(t)$, the time-dependent change of volume associated with the inflation and deflation of a magma chamber (see Figure 3). The maximum volume variation corresponds to ~4.9×10^4 mm^3, and occurs at ~4.3 yr.

In Section S2 and Table S1 of the Supplementary material all the parameters used for the different simulations are provided.

5.2 – Description of the sensor signals

In this work we attempt to reproduce only a few of the possible conditions that are known in the real world in terms of geodetic network geometry and data quality. In particular, we test 1) the possibility of different geometries of the GPS network around the fault source and 2) three different configurations of missing data in the synthetic time series.

GPS data include white and colored noise (Langbein, 2008), in particular pink (or flicker) and red (or Brownian or random-walk). We simulate the presence of the three sources of noise fixing them to pre-determined powers, resulting in an average total noise power over the 60 time series of ~1.1 mm^2. In order to vary the SNR we prefer to control the signal power through the source intensities, rather than changes the noise in the time series, since the noise generation in the synthetic time series is performed randomly. The two different geometries tested are shown in Figure 4, and we refer to them as Network 1 (N1) and Network 2 (N2). The possibility of a GPS network to record a particular source signal depends on the geometry of the network w.r.t. the position of the source and the source intensity. Here, we use the inter-seismic locking depth as a reference distance to define the geometric quality of a GPS network. In particular, N1 has half of the total number of stations (i.e., 10) located in a square box of side equal to twice the locking depth, and centered at the surface projection of the mid-point of the fault top edge. The network N2 consists in 20 stations, but only 1/10 of them are located in the box described above.

Real cGPS time series often present broad empty data gaps, mainly due to malfunctioning or theft of the geodetic instrumentation (antenna, batteries or receiver), rather than data missing com-

pletely at random. In order to create missing data in the synthetic time series we randomly create a bunch of gaps with different lengths, and delete the data associated with these gaps, obtaining two dataset structures (i.e., masks) with 5% and 25% missing data, respectively. These two masks are applied to the datasets obtained modifying both the source intensities and the network configurations. We do this in order to avoid to continuously generate random gaps, that would be again a factor out of control during the analysis.

5.3 – Case studies

Figure 5 summarizes the case studies we consider. The generation of the synthetic dataset results from the following steps:

1) Definition of the area extent under study;

2) Creation of the fault geometry and the creeping section;

3) Creation of the GPS network;

4) Creation of the source signals: inter-seismic, co-seismic, post-seismic, seasonal, Mogi;

5) Calculation of the surface displacements;

6) Addition of noise to the time series: white, flicker, and random walk noise;

7) Removal of missing data according to three pre-defined masks (0%, 5%, and 25% of missing data).

We assign to the estimated positions an uncertainty of 2 and 5 mm for the horizontal and vertical components, respectively.

It is necessary to point out that the co- and post-seismic sources described in Section 5.1 are not independent. Indeed, from the temporal evolution of the second source we can deduce something concerning the temporal evolution of the first one, and vice-versa. Let us imagine that we know the value of the post-seismic source at a certain epoch. If it is equal to 0 (or the reference point before the earthquake), then we know the value of the co-seismic source at the same time; if it is different from the pre-earthquake value, then again we know the value of the co-seismic source at

that epoch. It follows that the probabilities of the two sources do not factorize: $p(s^{co},s^{post})\neq p(s^{co})p(s^{post})$, but the application of ICA algorithms relies on such a factorization. Accordingly, we decide to pre-process the time series correcting them for the co-seismic jump, which is recovered using a PCA with the same number of components used for the subsequent ICA. In so doing, we assume that the PCA reproduces satisfactorily well the observed (simulated) time series data using few components, and we minimize the effect of the noise in the determination of the co-seismic offset. This, unfortunately, is not always the case, since sometimes we handle cases having a very low SNR (see Tables from S3 to S6 of the Supplementary material for the SNR values of each source and each configuration studied in this work). Moreover, also the geometric configuration of the network affects the SNR recorded. For the three co-seismic sources tested the percentage of time series with SNR > 1 is 70%, 45%, and 23% for the configuration N1, but these values drop to 56%, 10%, and 3% for the configuration N2. The same is valid for the post-seismic signal, passing from values of 63%, 28%, and 8% to values of 28%, 3%, and 0%. The linear and seasonal signals are less affected by the network configuration since they are less localized than the signals related to the seismic event. Indeed, for the linear signals tested this percentage is about 5%, 60%, and 95% for the 2 mm/yr, 12 mm/yr, and 60 mm/yr creep rates adopted, respectively, while for the seasonal signal it is 28-30% for both the network geometry configurations. Finally, for the case with also a Mogi source, the SNR related to the volcanic signal is greater than 1 for 6% of the stations.

The total number of cases studied is 55, coming from the use of 2 network configurations × 3 missing data masks × 3 earthquake energy release × 3 inter-seismic slip rates + 1 with a Mogi source (not reported in Figure 5). As an example, Figures 6a and 6b show the synthetic position time series of the stations 7 and 8 relative to the case where also the Mogi source is active, i.e. with a network configuration N2 (see Figure 4b for the location of the stations on the map). Figures 6c and 6d instead show the position time series of the stations 10 and 17 relative to the case with a network configuration N1 (see Figure 4a for the location of the stations on the map), 5% of missing data, a medium size earthquake, and a long term fault slip-rate of 2 mm/yr (i.e., slow tectonic rate).

6 – Results of PCA and ICA application on synthetic time series

In order to compare the results recovered from different algorithms on different datasets (obtained from different network configurations and tectonic regimes) we use the following quantities: the χ^2 and the Mean Squared Error (MSE) for the reconstructed sources. The χ^2 is defined as:

$$\chi^2 = \sum_{j=1}^{M}\sum_{t=1}^{T} o_{jt} w_{jt} (x_{jt} - x_{jt}^{decomp})^2 \tag{8}$$

where o_{jt} is the data mask (i.e., o_{jt} is equal to 0 if the data is missing and 1 if the data is recorded), w_{jt} is the weight associated to the data corresponding to the t-th epoch of the j-th time series, and it is equal to the inverse of the square of the uncertainty associated to the given data, x_{jt} is the actual data, and x_{jt}^{decomp} corresponds to the data as reconstructed by a given decomposition technique (PCA or ICA). The MSE is defined as:

$$MSE = \frac{1}{LT}\sum_{i=1}^{L}\sum_{t=1}^{T} (v_{it}^{true} - v_{it}^{decomp})^2 \tag{9}$$

where v_{it}^{true} is the value of the actual i-th source at time t, and v_{it}^{decomp} is the value of the i-th component at time t. L, M, and T have the same meaning as in Section 2.

The goal of the PCA is to minimize the χ^2. Instead, an ICA performs better in reconstructing the true sources (in the case when they are statistically independent), but provides higher χ^2 values than PCA. The lowest the value of the χ^2 and MSE quantities, the better is the fit to the data and the unravel of the sources, respectively. These features are testified by Tables 1 and 2, which show the percentage of improvement in the χ^2 using a PCA instead of an ICA ($\gamma_{\chi 2}$), and the percentage of improvement in the MSE using an ICA instead of a PCA (γ_{MSE}) for the low tectonic rate (2 mm/yr) scenario. These two quantities are defined as follows:

$$\gamma_{\chi^2} = \frac{\chi^2_{ICA} - \chi^2_{PCA}}{\chi^2_{PCA}}, \qquad \gamma_{MSE} = \frac{MSE_{PCA} - MSE_{ICA}}{MSE_{ICA}} \tag{10}$$

6.1 – Low tectonic rate (2 mm/yr)

For this scenario we consider 18 cases, using four synthetic sources (linear, co-seismic, post-seismic, and seasonal) and estimating the co-seismic one with a PCA (see section 5.3). If we use a reduced χ^2 test in order to select the number of components, we find that only one component is sufficient to explain the data, for both the PCA and the ICA decompositions. This choice is clearly incorrect, consequently, it is necessary to use a different test to select the number of components. Using an F-test for the PCA decomposition (following Kositsky and Avouac, 2010) we find that 3 components are not enough. This means that PCA introduces some noise into the reconstruction probably because the signals under investigation have a SNR < 1 in several cases (see Tables S3-S6 in the Supplementary Material). Instead, an F-test performed on the vbICA decomposition suggests that 2 components are sufficient in all cases but the one with 0% of data missing, the network configuration N1, and a co-seismic source corresponding to a $M_w = 6.85$ earthquake. In such a case, the F-test points out that the 3 components decomposition is the most appropriate one. Finally, the use of the criterion based on the ARD method suggests 2 components as the most appropriate 14 times over 18, and in the remaining 4 cases it suggests 3 components. In most of the cases (14 over 18) the linear signal is not properly recognized due to its low SNR (< 0.3, see Table S3, first three columns, in the Supplementary material). The criterion based on the ARD method suggests the use of 3 components only when 25% of the data is missing, but in all these cases one of the 3 ICs contains both the linear signal and a residual co-seismic offset. This means that the PCA reconstruction of the offset, and its consequent subtraction from the dataset, is not good enough, leaving in the processed data a residual offset. However in general this subtraction is necessary, as we will see at the end of this subsection. A similar problem affects also the two cases relative to a small earthquake scenario ($M_w^{co} = 5.94$ and $M_w^{ps} = 5.80$). In those cases where only two ICs are identified as necessary to explain the data, despite the use of a good network configuration (N1) and a small percentage of missing data, in one of the two ICs the post-seismic signal is superimposed to the linear trend (see Figures S2f and S3f of the Supplementary material). Moreover, the smaller the post-seismic

source, the more significant the linear signal contribution on the corresponding IC is (see Figures S2-S7 of the Supplementary material). The superposition of the linear trend and the post-seismic signal in the same IC might introduce some error in the estimate of the time decay constant of the latter signal, as we will discuss in the next Section.

As an example, Figure 7 compares the PCs and the ICs relative to the mid-size earthquake with 5% of missing data in the synthetic time series. The benefits of the ICA are evident if we look at Figures 7a and 7b, i.e. when we have a good network configuration (such as N1). We are not able to correctly infer the post-seismic signal in the cases when the signal intensity is low (M_w^{ps} = 5.80 and 6.09) and the cGPS network configuration is unfavorable (N2) (see Figure 7d and Figures S2-S7 of the Supplementary material). However in all cases but one, the vbICA algorithm reduces the MSE, as expected (Table 2). The only scenario where the MSE is better for a PCA is the worst possible case studied, i.e. the one with 25% of missing data, a small post-seismic source, and the N2 network configuration. For this configuration, the SNR of the post-seismic source is equal to 0.009, resulting in a very noisy component. Finally, as an example, we present in Figure 8 the results of the PCA and the vbICA decompositions performed on the position time series relative to the mid-size earthquake with 5% of missing data and a good network configuration (same as Figures 7a and 7b), but where the co-seismic offset is not corrected. The ARD method for the selection of the number of components coherently indicates that we should keep 3 ICs, that is one more w.r.t. the case where we have removed the co-seismic signal. As expected, even if we are using the network N1, the vbICA algorithm correctly isolates the seasonal signal, but does not succeed in separating the co- and post-seismic sources (see Figure 7b for comparison).

6.2 – High tectonic rate (12 and 60 mm/yr)

For high values of the inter-seismic slip rate the linear signal becomes dominant, and the entire dataset is strongly correlated. In other words, the cloud of points in the data space is strongly aligned, and this makes the search of the IC directions difficult. Choudrey (2002) pointed out that a correlation value greater than 0.67 prevents the vbICA to work properly. A possible solution, which

allows to maintain the linear signal in the time series (thus to avoid, for example, estimates of a linear trend in case of short time-series or in the presence of strong non-linear motions), consists in changing the reference frame into a station-fixed reference frame. Of course, the estimation of the linear trend can be performed only if we have at our disposal enough consecutive data for which the linear signal is the only non-stationary one, so that the estimation of the linear signal is not significantly affected by the other superimposed signals (e.g., Blewitt and Lavallée, 2002). For the synthetic data under study we estimate the linear trend using the 3 years of observations before the earthquake, i.e. we do not consider post-seismic data in order to prevent a distorted estimation for the trend. Such a solution is helping only for the network configuration N2. This is due to the fact that we use as a reference station the station number 1 (see Figure 4). In the N2 case most of the stations are located North-West of the source, close to the reference station. It follows that the linear trend is largely reduced, and the correlation of the dataset is reduced below the 0.67 threshold mentioned above.

Among all the N2 cases, we have already proven that in the case of a low tectonic rate the post-seismic signal is not recovered for the M_w^{ps} = 5.80 and 6.09 scenarios because of the very small SNR. All the more this is true at high tectonic rates, since the SSR between the post-seismic and the linear signals is even lower, and the transient signal of interest is undetectable. In Figure 9 we show that in the case of 0% of missing data and a large earthquake the vbICA decomposition reduces the global MSE value, but a strong cross-talk between the ICs representing the linear and the post-seismic sources is still present (Figure 9b). This is due to the fact that the vbICA algorithm is trapped near a local maximum for the NFE, close to the PCA solution. In order to allow the algorithm to escape from it, it is possible to try different random initializations. Figure 9c shows the ICA sources corresponding to the random initialization which gives the highest NFE among ten random initializations.

6.3 – Mogi source

The number of sources used for this synthetic test is four: a co-seismic Heaviside step function, a logarithmic post-seismic decay function, a seasonal sinusoidal function, and a linear combination of arctangent functions for the volcanic source (see Figure 3 and Section S2.2 of the Supplementary material). Due to the location of the Mogi source and the relatively small spatial extent of the region affected by its presence, we test only the N2 network configuration. We have already shown that this network configuration is able to properly recover the co- and post-seismic signals only for the seismic scenario corresponding to $M_w^{co} = 6.85$, $M_w^{post} = 6.57$. Consequently, we generate the data using only such a scenario, using the 5% missing data mask. After the correction of the co-seismic step, the criterion based on the ARD method correctly suggests to use 3 components. The results of a PCA and a vbICA are shown in Figure 10a and 10c. The loss in the χ^2 passing from the PCA to the vbICA is $\gamma_{\chi 2}$~3%, while the gain in the MSE, γ_{MSE}, is greater than 500%, as it can be visually seen looking at Figures 10a and 10c.

7 – Discussions

The tests performed suggest that the configuration of the geodetic network w.r.t. the source is critical. A well designed geodetic network is mandatory if we want to detect signals with small intensities (SNR < 1). In critical situations (i.e., less than 5% of the stations show a SNR > 1, and those few stations do not have a relevant SSR compared to other signals) the performance of the multivariate statistical techniques we have tested is poor, mainly because the signal is too localized and not common to a sufficient number of stations in the network. Unfortunately, in real data applications we do not know the actual SNR value, and we have to figure out which are the relevant signals recorded by the data. We have shown that having 10 cGPS stations within a box of side equal to two times the locking depth of a fault where an earthquake of $M_w > 6$ occurred allows us to solve satisfactorily the BSS problem with the proposed algorithm. In particular, for slow long-term slip rates (2 mm/yr, Section 6.1), a 99.99% credible interval contains the correct value of the used temporal decay constant τ in 12 over 18 cases (see Section S3 of the Supplementary material for the derivation of the credible intervals). This finding indicates that, even if in 14 over 18 cases we use

only 2 ICs to recover 3 sources (see Section 6.1), in most of the cases we are able to properly infer a correct credible interval for the temporal decay parameter (τ). This is probably due to the fact that the potential corruption of the post-seismic IC due to the residual linear signal is negligible. In the remaining 6 cases the temporal decay constant parameter is not resolved, and all the values spanned are acceptable or the post-seismic source has not been identified. These 6 cases correspond to the N2 geometry scenarios for the intermediate and small post-seismic source intensities. This proves that it is necessary to have a good quality geodetic network if we want to study crustal displacements of the order of few mm with multivariate statistical techniques such as vbICA.

Among all the tested scenarios, those with a high tectonic rate (> 10 mm/yr) show a highly correlated (> 0.67) dataset. Such a high correlation may prevent the vbICA algorithm to identify the proper directions in the data space (i.e., the independent components) onto which to project the data. The removal of a linear trend from the original time series reduces the correlation of the dataset. In section 6.2 we subtracted to all the stations the corresponding horizontal, and eventually also the vertical, long term velocities derived from a given reference station. This strategy provides satisfactory results if the reference station is located in the proximity of the available network. For example, in Figure 9c we show the case of 0% of missing data, a large earthquake, an unfavorable network (N2), and a tectonic rate of 60 mm/yr. The credible interval at 99.99% for the post-seismic decay constant correctly includes the 1 day temporal decay used for the logarithmic source. An alternative solution to the high correlation issue may be to refer the time series data to an external stable reference frame (e.g., a plate-fixed frame). However, this strategy implies to consider the linear trend as an actual signal, and it may distort the identification of signals having a curvature different from zero (e.g., acceleration of the ground due to isostatic adjustment). It is also possible, as proposed by Choudrey (2002), to decorrelate the dataset from the very beginning, for example using a PCA. However, this procedure is likely to wipe out signals of potential interest.

In absence of missing data the BSS problem is over-determined, i.e. the number of sources L is lower than or equal to the number of time series M. If it happens that only n stations recorded

their position in a given time interval, then the maximum number of components that we can estimate in the same time interval is $3n$. The problem of missing data is common to geophysical data in general, and it is relevant in GPS measurements, where daily position data are usually not missing completely at random. Looking at the ability of the vbICA algorithm to recover the original sources (see Table 2, bottom), we find that the sources are better recovered when the percentage of missing data is low. We did not perform systematic tests with more than 25% of missing data, but in few cases with ~50% of gaps we noticed a considerable degradation both in the PCA initialization and in the final vbICA solution. Even if PCA provides lower χ^2 values w.r.t. the ones computed from a vbICA reconstruction (i.e., positive values for $\gamma_{\chi 2}$, see Table 1, upper part), vbICA performs better than PCA with the estimate of missing data, as we explain in the following. Considering the low tectonic rate scenario, we reconstruct the position time series, $x_{jt}^{\ decomp}$, using the PCs and the ICs deduced from the decomposition of the dataset with two different percentages of missing data (5% and 25%). The reconstructed time series estimate the position even at the epochs when the data actually used for the same decomposition are missing (missing epochs). Then, particularly for those epochs, we can compare such a reconstruction with the actual data (known from the synthetic unmasked time series) and we find that the correspondent position values are better explained by the vbICA decomposition. Indeed, we evaluate the $\chi^{2\ \mathrm{MD}}$, calculated as in equation (8) but using $w_{jt}^{MD} = (1 - o_{jt})w_{jt}$ in place of $o_{jt}w_{jt}$, $j = 1, \ldots, M$ and $t = 1, \ldots, T$, where o_{jt} is the data mask. In 9 over the 12 cases reported in Table 1, $\chi^{2\ \mathrm{MD}}$ is lower if we use the vbICA reconstruction (see Table 1, bottom). We argue that such a better performance of vbICA w.r.t. PCA in reconstructing the missing data is due to the fact that it provides a more faithful representation of the actual sources, allowing a better prediction of the values at the missing epochs. We think that the more trustful filling of the data gaps obtained through the vbICA decomposition may be useful to better constraint geophysical model of the surface displacement.

The bayesian approach to the ICA brings other two advantages with respect to classic ICA and PCA. The first one consists in having the full posterior pdf of both the sources and the mixing ma-

trix. Such a knowledge allows us to propagate the uncertainties on the reconstructed time series, providing a more complete understanding of the actual reliability of the data decomposition. In this work we exploit this information also assessing the 99.99% credible interval related to the temporal decay constant parameter (τ) of the post-seismic signal. The second advantage allows us to handle the classical multivariate statistical methods' issue regarding the determination of the proper number of components that best describe the original data. Here the selection of the number of components is done using a criterion based on the ARD method and adopting loose prior parameters so that the data can reveal their inner structure (see Section 2). After having determined the proper number of components, it is necessary to tune the prior parameters, and this tuning can be performed evaluating the NFE function.

As well as for the original code of Choudrey (2002), also in a case of missing data the vbICA algorithm outperforms the FastICA one, which is based on a mapping approach (see Section 2). In the analysis of the superposition of a seasonal signal, a post-seismic decay, and a time-varying volcanic source, we have compared the ICA decomposition obtained using the FastICA and the vbICA algorithms (see Figures 10b and 10c). While the FastICA decomposition is not able to correctly separate the volcanic and post-seismic sources, the vbICA clearly separates the two contributions. In particular, the gain in the MSE using the FastICA w.r.t the PCA decomposition is ~44%, while using the vbICA gives us an improvement of more than 500% in the MSE. Despite the low percentage of stations with SNR > 1 (~6%), the signal of the volcanic source is correctly recovered (see Figure 10c). Moreover, the post-seismic decay constant is correctly inferred at a 99.99% credible interval from the first IC deduced from the vbICA algorithm. This better performance must be ascribed to the greater flexibility of the vbICA approach w.r.t. the more classic mapping approaches. Figure 10d shows that the pdfs corresponding to the three ICs of Figure 10c have a clear multimodal nature, which is completely captured using the mix of Gaussians of the vbICA approach.

Moreover the capability to recover the actual temporal sources enables vbICA to reconstruct a more faithful mixing matrix w.r.t. the one obtained with different multivariate statistical techniques,

like PCA or FastICA. This in turns allows us to visualize in a more clear way the spatial extent of the ongoing geophysical processes.

8 - Conclusions

Real world ground displacement observations are the result of a combination of different sources, and in order to reveal the inner structure of the data it is useful to search for an intrinsic coordinate frame where the informative content of the data is maximized. For this reason, we propose the use of an advanced ICA technique based on variational approximation and bayesian inference (vbICA), with the goal of solving the BSS problem in ground displacement time series, as those recorded by GPS networks, and extract the robust information about the sources originating the observations.

A modified vbICA algorithm, which allows to handle missing data that are a common problem in GPS observations, has been tested on synthetic time series generated in order to simulate earthquake and volcanic deformations. We find that the vbICA method shows clear advantages w.r.t. widely used multivariate statistical techniques like classic PCA or ICA in extracting the relevant information from high-dimensional (>3) datasets, such as the ones typically studied in tectonic geodesy. The main benefits derived from the vbICA consist in: 1) a better characterization of the sources, which leads also to a better forecast of the missing values as testified by the lower $\chi^{2\ MD}$; 2) a full description of the sources' pdf, which allows us to estimate also the uncertainty related to the model parameters inferred from the ICs.

On the basis of our tests, we suggest to use the vbICA technique adopting the following steps:

1) Center the dataset, i.e. remove the mean to each time series

2) Check the correlation of the centered dataset

2a) if the correlation is greater than 0.67, go to point 3)

2b) if the correlation is smaller than 0.67, go to point 4)

3) Detrend the time series

4) Correct for the co-seismic offsets because of the non independence with the post-seismic signals

5) Perform a vbICA with loose priors and select the number of components via a criterion based on the ARD method

The vbICA is performed starting from a PCA initialization. If the ICs are very close to the PCs it is possible that the algorithm stucks in a local maximum for the NFE, corresponding to the PCA solution. In this case we find that a random initialization helps retrieving the original sources from the data.

The method and approaches discussed in this work can be applied to the analysis of any kind of geodetic time series data, including for example InSAR, strainmeters, and tiltmeters. In particular the results presented in this work show the capability of a vbICA analysis in detecting transient deformations in a spatially and temporally consistent way.

Acknowledgements

The displacements at the surface relative to the Mogi source have been calculated using the code of François Beauducel (http://www.ipgp.fr/~beaudu/matlab.html#Mogi, last access 10 March 2015). This study has benefited from funding provided by the Italian Presidenza del Consiglio dei Ministri – Dipartimento della Protezione Civile (DPC). This paper does not necessarily represent DPC official opinion and policies. Comments from the Editor-in-Chief, Jürgen Kusche, the Associate Editor, Danan Dong, and three anonymous reviewers improved the paper.

Figure captions

Figure 1: Typical geophysical signals commonly determined in GPS position time series, and histograms of the corresponding pdfs: a) linear signal, b) seasonal signal, c) logarithmic decay signal, and d) flat pre-seismic + logarithmic post-seismic signal. Each panel is subdivided in two plots: on the right there is the temporal evolution of a specific signal, and on the left side there is the ten bins histogram showing the number of points belonging to a given bin. These histograms correspond to a rough approximation of the pdfs related to the corresponding signals. All the distributions are clearly non Gaussian, with a) showing a uniform distribution, b) showing a bimodal symmetric distribution, c) showing a unimodal asymmetric distribution, and d) showing a bimodal asymmetric distribution.

Figure 2: Fault model used to simulate earthquake cycle deformation recorded at a network of geodetic points at the surface. The green line corresponds to the intersection between the ground and the extension of the fault plane. For all the simulations we have used a rake of -90° (thrust regime). a) Slip distribution after 5 years. The yellow region indicates the patches that undergo only afterslip (200 mm); the light red region experiences only the co-seismic slip (400 mm); in the darker region there is a superposition of the two slip sources (total slip: 600 mm). b) As in case a), there are three different regions: a light yellow one (only afterslip), a light red one (only co-seismic slip) and a dark red one (co-seismic slip and afterslip). c) As a) and b), but with reduced intensities.

Figure 3: Temporal evolution of the Mogi source used in synthetic tests. The maximum displacement associated to the volcanic source recorded at one of the stations is ~13 mm.

Figure 4: GPS network configurations. Red triangles: cGPS stations. Numbers: station names. Green line: surface projection of the upper bound of the fault plane (see Figure 2). Black rectangles: surface projection of the fault patches. a) N1: Half of the total number of stations (i.e., 10) located in a square box of side equal to twice the locking depth, and centered at the surface projection of the mid-point of the upper bound of the fault plane. b) N2: As N1, but with only 1/10 of the total number of stations located in the box described above. Green circle: Mogi source (Figure 3) location.

Figure 5: Logic tree used for the creation of the synthetic data. After setting the tectonic regime (thrust faulting) and the time span (5 years), we study two different GPS network geometry configurations shown in Figure 4, as described in the main text. Then, we take into account three different masks for missing data (MD): 0%, 5%, and 25%. We consider three possible seismic sources which are characterized by a specific energy released through co-seismic and post-seismic slip: M_w^{cs} and M_w^{ps}. Finally, other three cases are considered, varying the slip rate of the creeping section of the fault in the inter-seismic phase. The addition of the noise is the same for all the data generated.

Figure 6: a) Position time series (black dots) recorded by station 7 relative to the case including a Mogi source (i.e., N2 network configuration, Figure 4b). b) As a), but station 8 is shown. c) Position time series (black dots) recorded by station 10 relative to the case with 5% of missing data, a medium size earthquake (M_w^{ps} = 6.09), low tectonic rate $(\dot{s}^{lin} = 2 \text{ mm/yr})$, and N1 network configuration (Figure 4a). d) As c), but station 17 is shown. The gray lines indicate the uncertainty related to each measurement.

Figure 7: Temporal evolution of the recovered PCs (a and c) and ICs (b and d) (black dots) in the case of 5% of missing data, medium earthquake scenario (M_w^{co} = 6.29 and M_w^{ps} = 6.09), and low tectonic rate$(\dot{s}^{lin} = 2 \text{ mm/yr})$.The gray lines in the ICs corresponds to the associated uncertainty related to the ICs, calculated as the square root of the variance. This estimation is enabled by the knowledge of the approximated pdf of each IC via a mix of Gaussians. The red lines correspond to the post-seismic and seasonal actual sources. a) and b): N1 cGPS network geometry. c) and d): N2 cGPS network geometry.

Figure 8: Left: PCs relative to the case MD 5%, N1, medium size earthquake, long term velocity 2 mm/yr not corrected for the co-seismic offset. Right: ICs for the same case. These decompositions should be compared with Figures 7a and 7b, where we have preprocessed the data correcting for the offset.

Figure 9: Temporal evolution of the recovered components (black dots) in the case of 0% of missing data, N2 cGPS network configuration (Figure 4b), large earthquake scenario (M_w^{co} = 6.85 and

M_w^{ps} = 6.57), and high tectonic rate ($\dot{s}^{lin}$ = 60 mm/yr). a) PCs. b) ICs using the PCA as a starting initialization for the vbICA algorithm. c) ICs using a random initialization for the vbICA algorithm. The gray lines in b) and c) are, as in Figure 7, the uncertainties associated to the ICs. The red lines correspond to the linear, post-seismic, and seasonal actual sources. The decomposition shown in c) corresponds to the one showing the best NFE value among the 10 random initializations tested. The random initialization allows the vbICA algorithm to escape from the local maximum of the PCA decomposition. The percentage MSE gain is ~100% for the PCA initialization, and it is ~6750% for the random initialization. The percentage χ^2 loss, instead, is ~2% and ~3%, respectively.

Figure 10: Decomposition results relative to the N2 cGPS network configuration (see Figure 4b) with 5% of missing data, and the mix of the M_w^{co} = 6.85 and M_w^{ps} = 6.57 co- and post-seismic source, the seasonal source, and the Mogi source (see Section 3 for further details and Figure 3). a) PCA decomposition. b) ICA decomposition with the FastICA algorithm (Hyvärinen and Oja, 1997). c) ICA decomposition with the modified vbICA algorithm tested in this work. Red lines in c) indicate the actual sources. d) Probability density functions associated to the ICs shown in c). The continuous black lines correspond to the sum of the mix of Gaussian used to mimic the pdf of the sources. We have used 4 Gaussian for each source, and they are indicated by the dashed colored lines.

References

Anderson E.M. (1936) Dynamics of the formation of cone-sheets, ring-dykes, and cauldron-subsidences, Proc. R. Soc. Edinburgh, 56, 128-157.

Argus D.F., Fu Y. and Landerer F.W. (2014) Seasonal variation in total water storage in California inferred from GPS observations of vertical land motion. Geophys. Res. Lett., Vol. 41, Issue 6, pp. 1971-1980, DOI: 10.1002/2014GL059570.

Barbot S. and Fialko Y. (2010) A unified continuum representation of post- seismic relaxation mechanisms: semi-analytic models of afterslip, poroelastic rebound and viscoelastic flow. Geophys. J. Int., 182, 1124–1140.

Blewitt G. and Lavallée D. (2002) Effect of annual signals on geodetic velocity. J. Geophys. Res. Solid Earth, Vol. 107, Issue B7, DOI: 10.1029/2001JB000570.

Bottiglieri M., Falanga M., Tammaro U., Obrizzo F., De Martino P., Godano C. and Pingue F. (2007) Independent component analysis as a tool for ground deformation analysis. Geophys. J. Int., 168, 1305-1310, doi: 10.1111/j.1365-246X.2006.03264.x.

Cardoso J.F. and Souloumiac A. (1993) Blind beamforming for non-Gaussian signals. Radar and Signal Processing, IEE Proceedings F, vol. 140, issue 6, pp. 362-370.

Chan K., Lee T.-W. and Sejnowski T.J. (2003) Variational Bayesian Learning of ICA with Missing Data. Neural Comput., 15(8):1991–2011.

Choudrey R.A. (2002) Variational Methods for Bayesian Independent Component Analysis. Pattern analysis and machine learning - robotics research group, University of Oxford. [Available at http://www.robots.ox.ac.uk/~parg/projects/ica/riz/thesis.html, last access 10 June 2015]

Comon P. (1994) Independent Component Analysis: A new concept? Signal Processing, **36**, 287-314.

Dong D., Fang P., Bock Y., Webb F., Prawirodirdjo L., Kedar S. and Jamason P. (2006) Spatiotemporal filtering using principal component analysis and Karhunen-Loeve expansion approaches

for regional GPS network analysis. J. of Geophys. Res., VOL. **111**, B03405, doi: 10.1029/2005JB003806.

Forootan E. and Kusche J. (2012) Separation of global time-variable gravity signals into maximally independent components. J. Geodesy, Vol. 86, Issue 7, pp. 477-497.

Forootan E. and Kusche J. (2013) Separation of deterministic signals using independent component analysis (ICA). Stud. Geophys. Geod., Vol. 57, Issue 1, pp. 17-26.

Heki K., Miyazaki S. and Tsuji H. (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. Nature, VOL. 386, 10 April.

Hyvärinen A. and Oja E. (1997) A Fast Fixed-Point Algorithm for Independent Component Analysis. Neural Comput., Vol. 9, No. 7, pp. 1483-1492, doi:10.1162/neco.1997.9.7.1483.

Ito Y., Hino R., Kido M., Fujimoto H., Osada Y., Inazu D., Ohta Y., Iinuma T., Ohzono M., Miura S., Mishina M., Suzuki K., Tsuji T. and Ashi J. (2013) Episodic slow slip events in the Japan subduction zone before the 2011 Tohoku-Oki earthquake. Tectonophysics, 600:14 26, doi:10.1016/j.tecto.2012.08.022.

Ji K.H. and Herring T.A. (2011) Transient signal detection using GPS measurement: Transient inflation at Akutan volcano, Alaska, during early 2008. Geophys. Res. Lett., Vol. 38, L06307, doi:10.1029/2011GL046904.

Johanson I.A., Fielding E.J., Rolandone F. and Burgmann R. (2006) Coseismic and Postseismic Slip of the 2004 Parkfield Earthquake from Space-Geodetic Data. Bull. Seismol. Soc. Am., vol. 96, no. 4B, S269-282, doi: 10.1785/0120050818.

Kositsky A.P. and Avouac J.P. (2010) Inverting geodetic time series with a principal component analysis-based inversion method. J. Geophys. Res., VOL. **115**, B03401, doi: 10.1029/2009JB006535.

Langbein J. (2008) Noise in GPS displacement measurements from Southern California and Southern Nevada. J. Geophys. Res., VOL. 113, B05405, doi:10.1029/2007JB005247.

Lohman R.B. and Murray J.R. (2013) The SCEC Geodetic Transient-Detection Validation Exercise. Seismol. Res. Lett., Vol. 84, no. 3, p. 419-425, doi: 10.1785/0220130041.

MacKay D.J.C. (1994) Bayesian non-linear modelling for the prediction competition. ASHRAE Trans., 100(2), 1053-1062.

Marone C., Scholtz C.H. and Bilham R. (1991) On the Mechanics of Earthquake Afterslip. J. Geophys. Res., VOL. 96, NO. B5, pp. 8441-8452.

Mogi, K. (1958) Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. Bull. Earthquake Res. Inst., Univ. Tokyp, 36, 99-134.

Okada, Y. (1985) Surface deformation to shear and tensile faults in a half- space. Bull. Seismol. Soc. Am., 75, 1135–1154.

Ormerod J.T. and Wand M.P. (2010) Explaining Variational Approximations. Am. Stat., Vol. 64, No. 2, DOI: 10.1198/tast.2010.09058.

Owen S., Segall P., Lisowski M., Miklius A., Murray M., Bevis M. and Foster J. (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. Geophys. Res. Lett., VOL. **27**, No. 17, pages 2757-2760.

Perfettini H., Avouac J.P., Tavera H., Kositsky A., Nocquet J.M., Bondoux F., Chlieh M., Sladen A., Audin L., Farber D.L. and Soler P. (2010) Seismic and aseismic slip on the Central Peru megathrust. Nat. Lett., 465, doi:10.1038/nature09062.

Riel B., Simons M., Agram P. and Zhan Z. (2014) Detecting transient signals in geodetic time series using sparse estimation techniques. J. Geophys. Res. Solid Earth, 119, 5140-5160, doi:10.1002/2014JB011077.

Roberts S.J. and Everson R.. Indepdendent Component Analysis: Principles and Practice. Cambridge University Press, 2001.

Segall P. (2010) Earthquake and Volcano Deformation. Princeton University Press.

Srebro N. and T. Jaakkola (2003) Weighted low-rank approximations. In Twentieth International Conference on Machine Learning, 2003.

http://ttic.uchicago.edu/~nati/Publications/SrebroJaakkolaICML03.pdf.

Lohman R.B. and Murray J.R. (2013) The SCEC Geodetic Transient-Detection Validation Exercise. Seismol. Res. Lett., Vol. 84, no. 3, p. 419-425, doi: 10.1785/0220130041.

MacKay D.J.C. (1994) Bayesian non-linear modelling for the prediction competition. ASHRAE Trans., 100(2), 1053-1062.

Marone C., Scholtz C.H. and Bilham R. (1991) On the Mechanics of Earthquake Afterslip. J. Geophys. Res., VOL. 96, NO. B5, pp. 8441-8452.

Mogi, K. (1958) Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. Bull. Earthquake Res. Inst., Univ. Tokyp, 36, 99-134.

Okada, Y. (1985) Surface deformation to shear and tensile faults in a half- space. Bull. Seismol. Soc. Am., 75, 1135–1154.

Ormerod J.T. and Wand M.P. (2010) Explaining Variational Approximations. Am. Stat., Vol. 64, No. 2, DOI: 10.1198/tast.2010.09058.

Owen S., Segall P., Lisowski M., Miklius A., Murray M., Bevis M. and Foster J. (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. Geophys. Res. Lett., VOL. **27**, No. 17, pages 2757-2760.

Perfettini H., Avouac J.P., Tavera H., Kositsky A., Nocquet J.M., Bondoux F., Chlieh M., Sladen A., Audin L., Farber D.L. and Soler P. (2010) Seismic and aseismic slip on the Central Peru megathrust. Nat. Lett., 465, doi:10.1038/nature09062.

Riel B., Simons M., Agram P. and Zhan Z. (2014) Detecting transient signals in geodetic time series using sparse estimation techniques. J. Geophys. Res. Solid Earth, 119, 5140-5160, doi:10.1002/2014JB011077.

Roberts S.J. and Everson R.. Indepdendent Component Analysis: Principles and Practice. Cambridge University Press, 2001.

Segall P. (2010) Earthquake and Volcano Deformation. Princeton University Press.

Srebro N. and T. Jaakkola (2003) Weighted low-rank approximations. In Twentieth International Conference on Machine Learning, 2003.

http://ttic.uchicago.edu/~nati/Publications/SrebroJaakkolaICML03.pdf.

Figure1

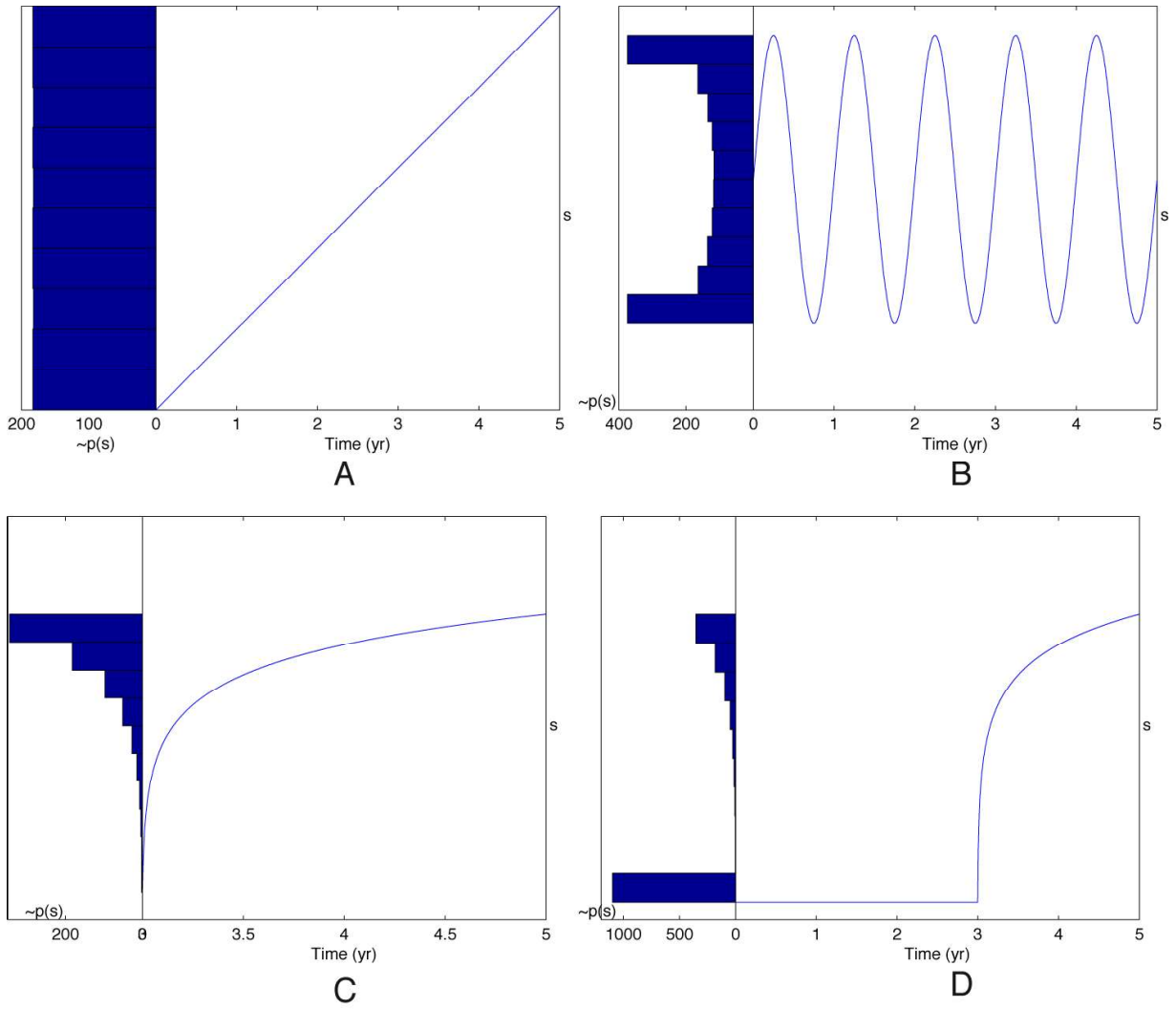


Figure 1

Figure2

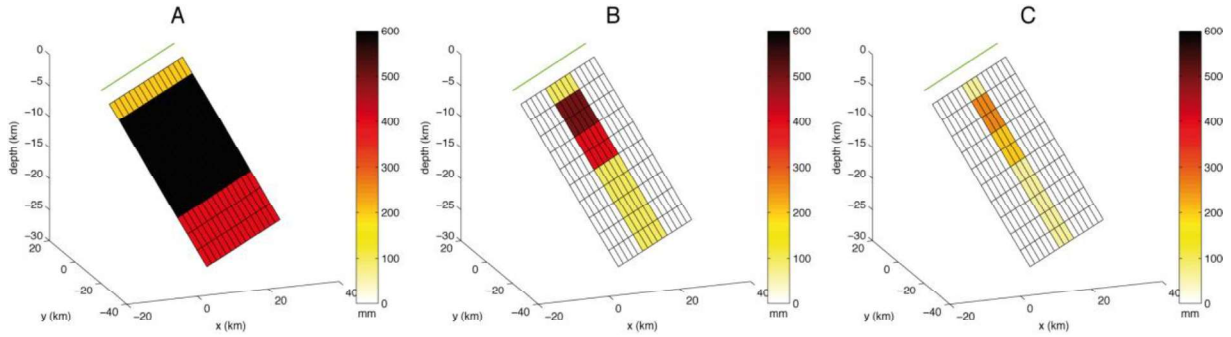


Figure 2

Figure3

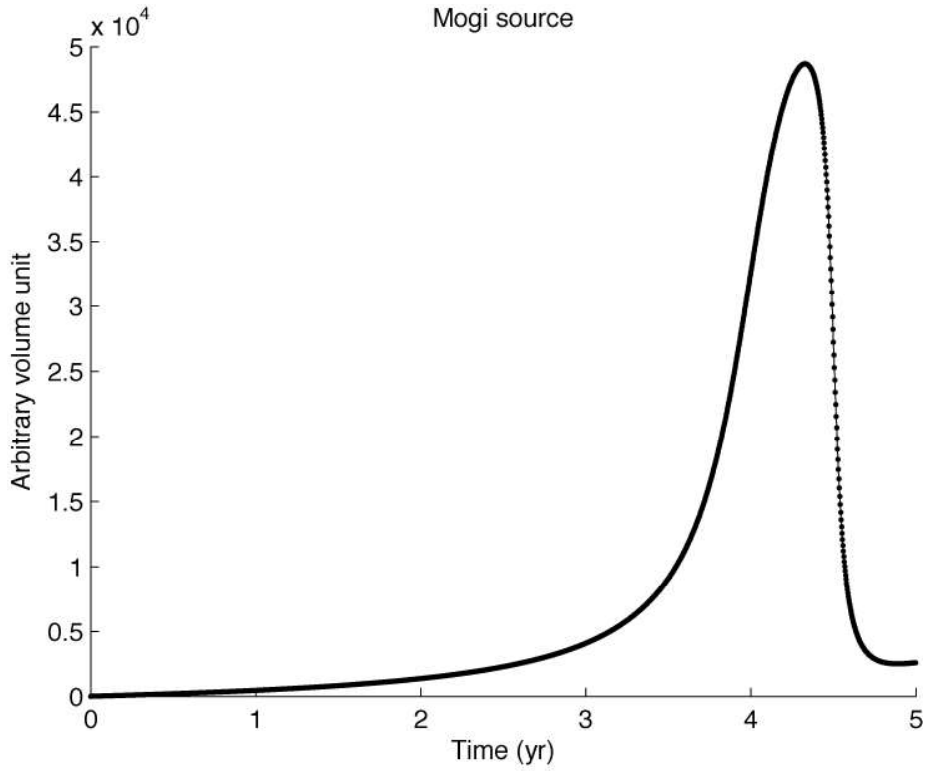


Figure 3

Figure4

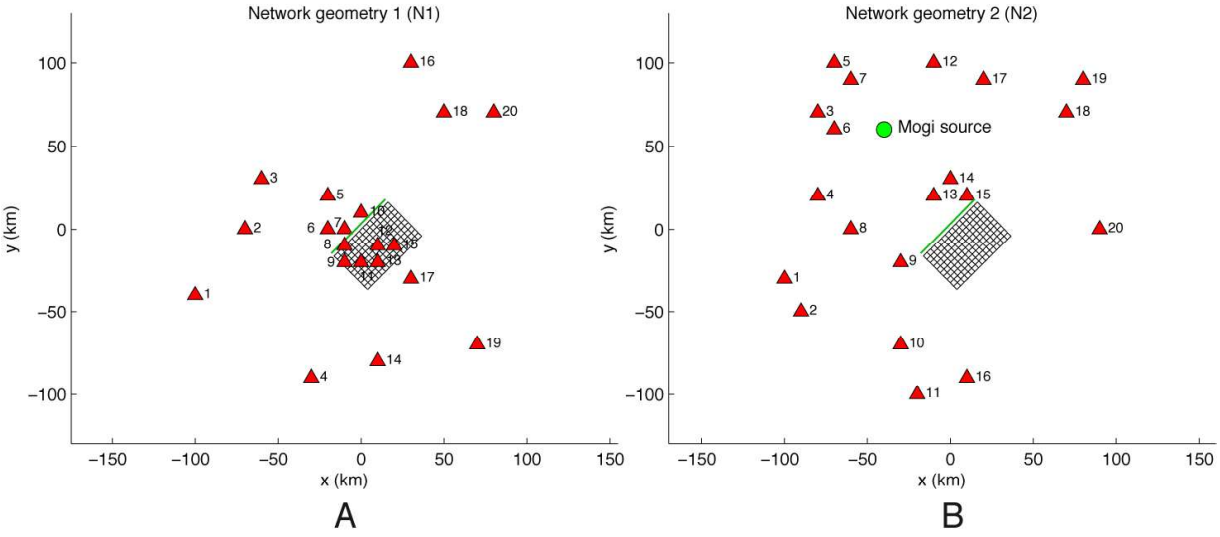


Figure 4

Figure5

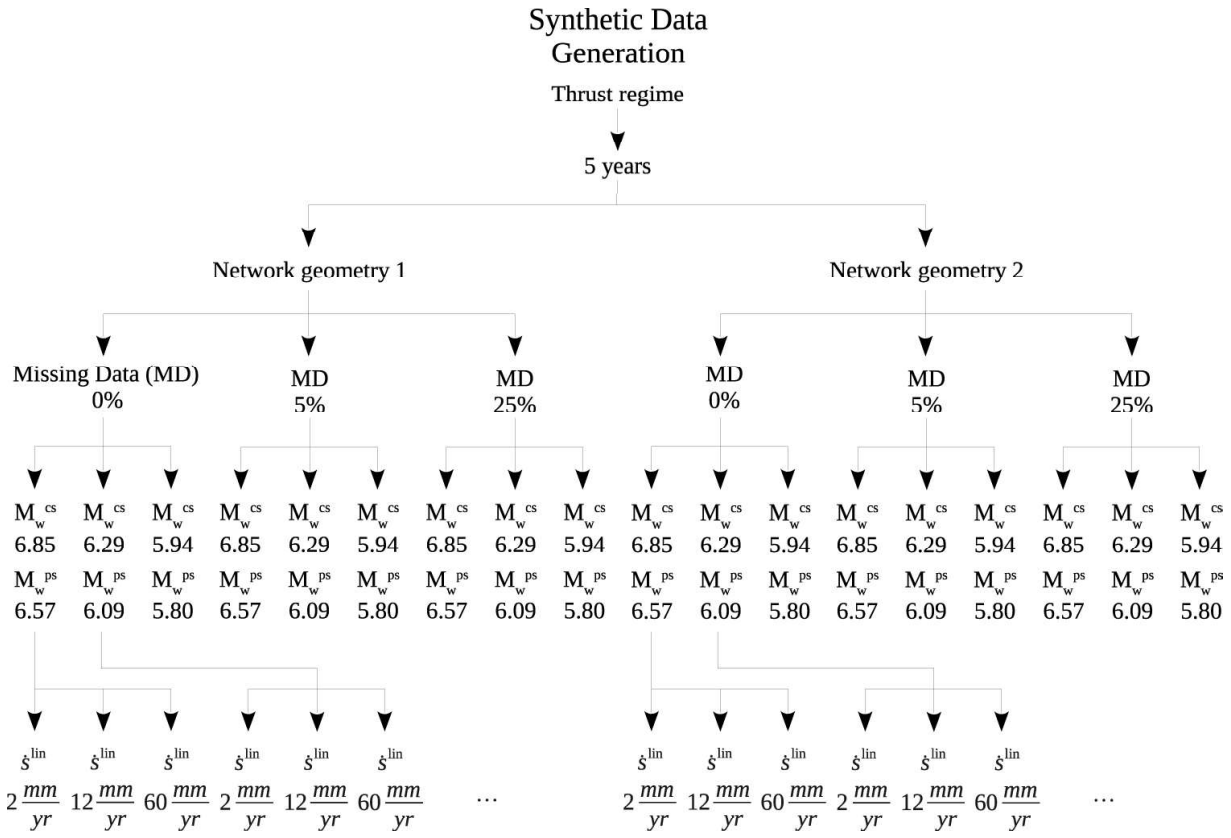


Figure 5

Figure6

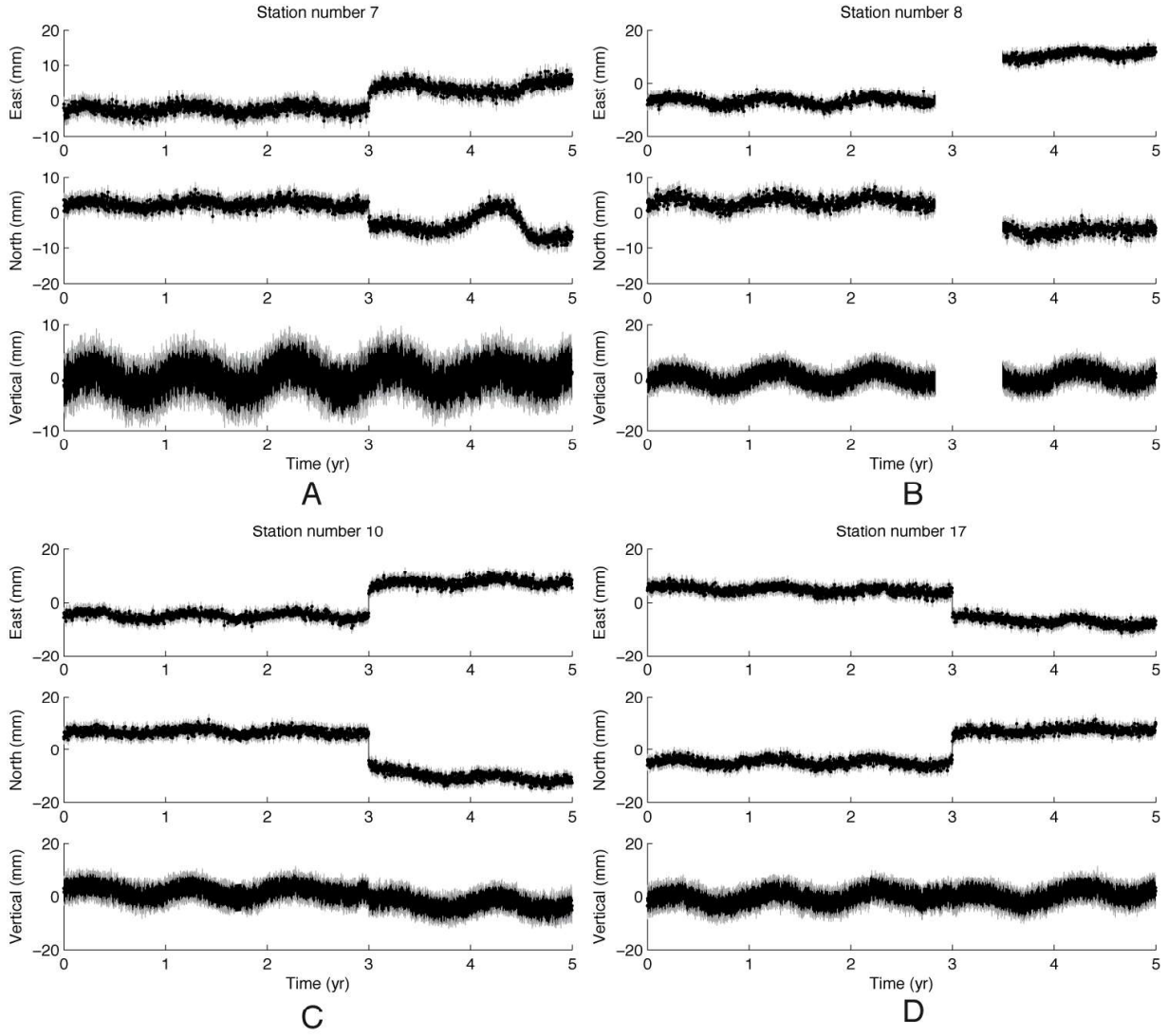


Figure 6

Figure7

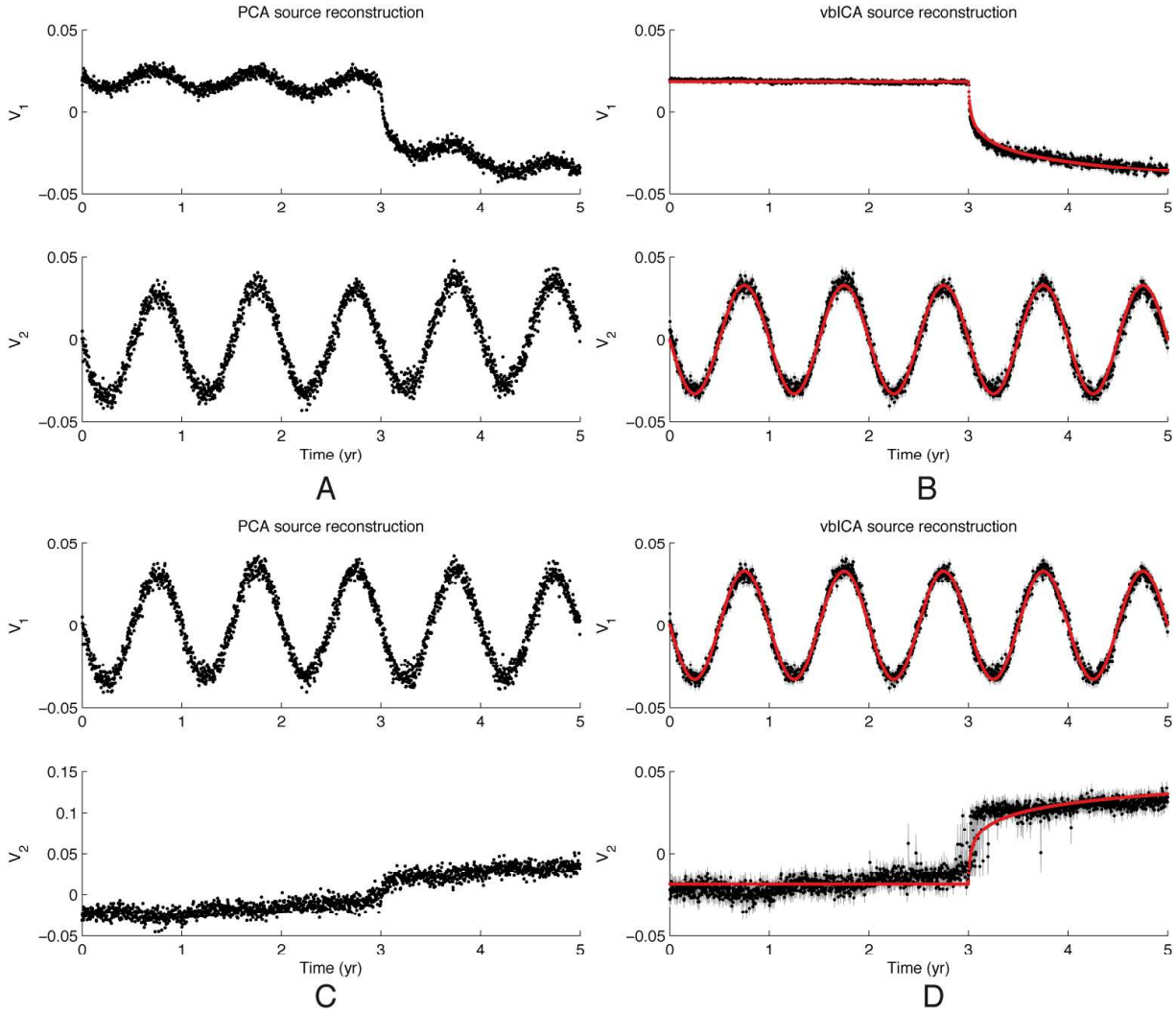


Figure 7

Figure8

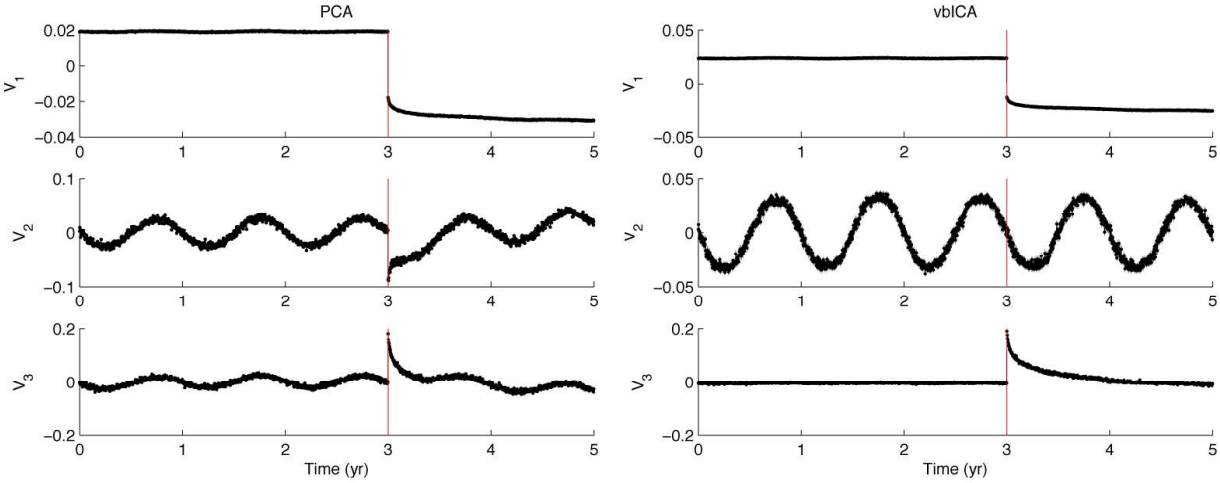


Figure 8

Figure9

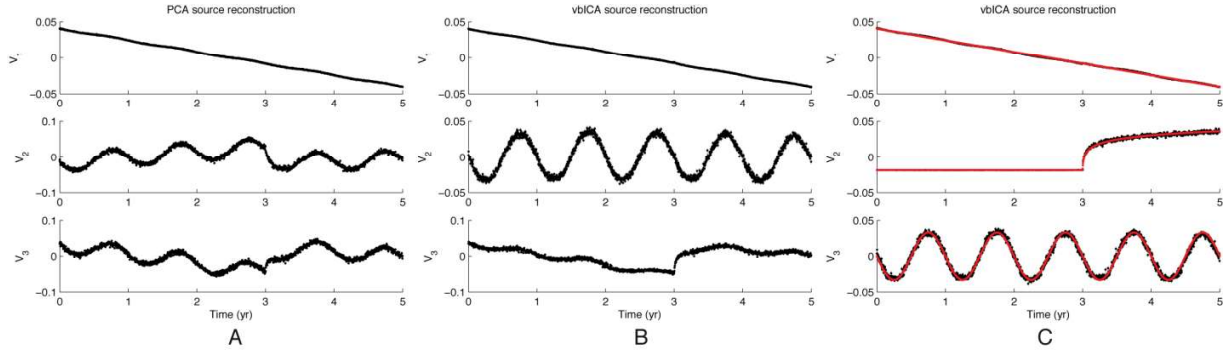


Figure 9

Figure10

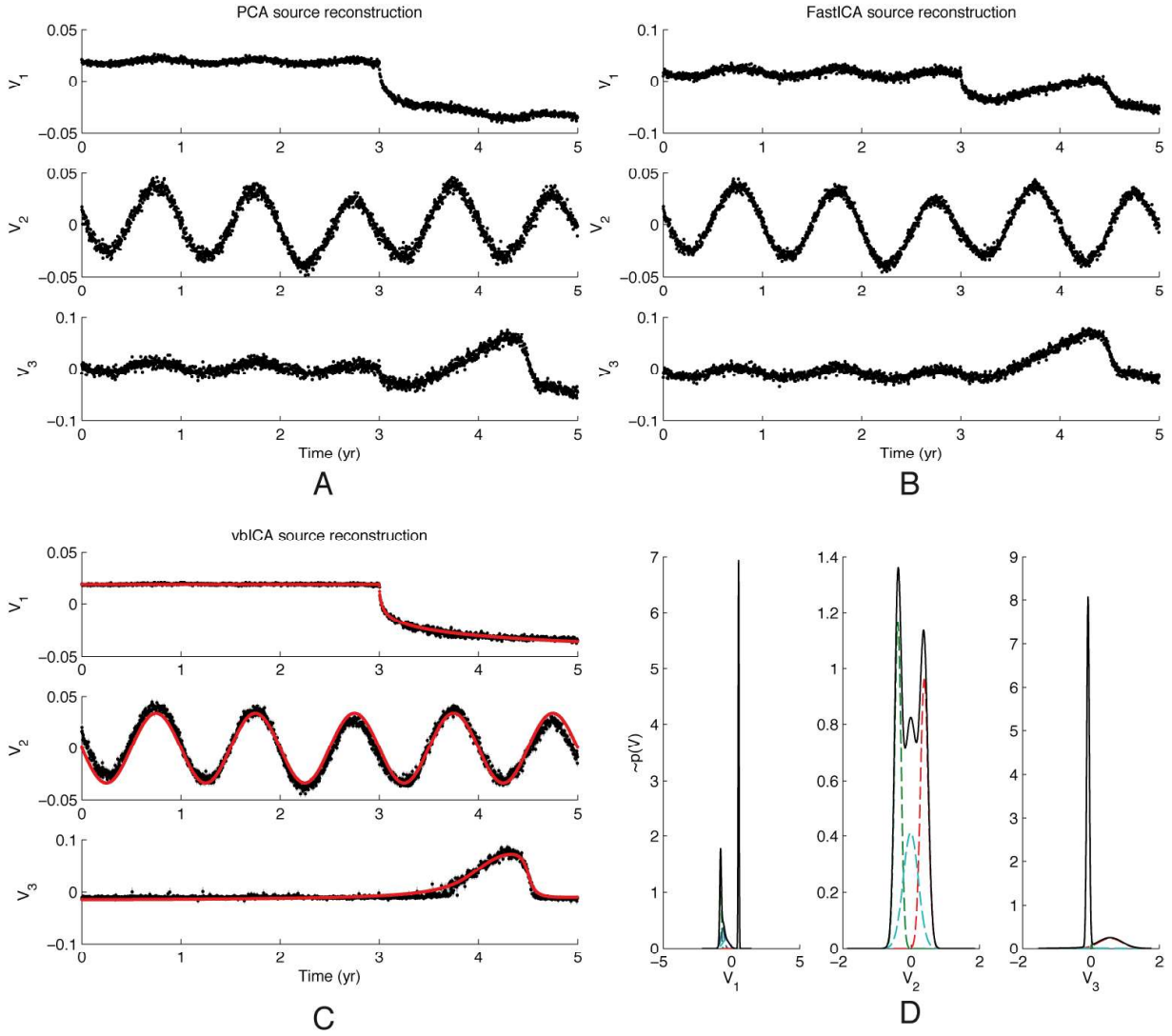


Figure 10

Tables

γ_{χ^2}		$\dot{s}^{lin}$ = 2mm/yr		
		MD 0%	MD 5%	MD 25%
M_w^{co} = 6.85	N1	4.8%	3.4%	5.2%
	N2	1.9%	1.0%	6.1%
M_w^{co} = 6.29	N1	2.3%	2.6%	5.5%
	N2	1.6%	1.9%	2.9%
M_w^{co} = 5.94	N1	2.4%	2.6%	7.7%
	N2	2.1%	2.2%	7.6%

$\chi^2_{decomp}{}^{MD} = \sum_{j=1}^{M}\sum_{t=1}^{T} w_{jt}^{MD}\,(x_{jt} - x_{jt}^{decomp})^2$		MD 5%		MD 25%	
		$\chi^2_{PCA}{}^{MD}$	$\chi^2_{ICA}{}^{MD}$	$\chi^2_{PCA}{}^{MD}$	$\chi^2_{ICA}{}^{MD}$
M_w^{co} = 6.85	N1	2944	3585	20918	28605
	N2	1150	1142	12395	11247
M_w^{co} = 6.29	N1	1149	1131	15103	11366
	N2	1070	1083	12587	12403
M_w^{co} = 5.94	N1	1105	1096	570537	130228
	N2	1053	1046	407138	91476

Table 1: Top: Percentage gained in the χ^2 quantity if the PCA decomposition is used instead of the ICA. M_w^{co} defines the seismic scenario (see Figure 5), N1 or N2 indicate the cGPS network configuration (see Figure 4), and MD stands for Missing Data. Bottom: $\chi^2_{PCA}{}^{MD}$ and $\chi^2_{ICA}{}^{MD}$ values, computed as in formula (8) but relative to all and only the epochs of missing data (for the purpose of the decomposition). The cases marked in green shows a decreasing value for the χ^2 of the missing data passing from the PCA to the vbICA decomposition.

γ_{MSE}		$\dot{s}^{lin} = 2$mm/yr		
		MD 0%	MD 5%	MD 25%
$M_w^{co} = 6.85$	N1	44%	49%	72%
	N2	221%	130%	33%
$M_w^{co} = 6.29$	N1	226%	230%	30%
	N2	76%	92%	12%
$M_w^{co} = 5.94$	N1	1767%	1430%	47%
	N2	25%	35%	-38%
MSE post+seasonal		MD 0%	MD 5%	MD 25%
$M_w^{co} = 6.85$	N1	0.0241	0.0245	0.0310
	N2	0.0187	0.0242	0.0365
$M_w^{co} = 6.29$	N1	0.0254	0.0261	0.0328
	N2	0.0910	0.0742	0.2755
$M_w^{co} = 5.94$	N1	0.0405	0.0430	0.1085
	N2	0.1523	0.1281	1.9872

Table 2: Top: Percentage gained in the MSE quantity if the ICA decomposition is used instead of the PCA. Bottom: MSE values for the post-seismic and seasonal sources. We notice a deterioration of the MSE with an increasing percentage of missing data (MD). Symbols as in Table 1.

Supporting Information for:

Blind Source Separation problem in GPS time series

submitted to Journal of Geodesy

A. Gualandi[1,3,], E. Serpelloni[2] and M.E. Belardinelli[3]*

1 Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna

2 Istituto Nazionale di Geofisica e Vulcanologia, Centro Nazionale Terremoti

3 Dipartimento di Fisica e Astronomia, Settore di Geofisica, Università di Bologna

* Corresponding author: Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Via D. Creti, 12, 40128 Bologna (IT). Phone: +39 051 4151474, e-mail: *adriano.gualandi@bo.ingv.it*

S1 – The variational bayesian ICA: some details

As said in Section 2 of the main text, under a bayesian framework the goal is to evaluate the quantity given by:

$$p(\mathbf{W}|\mathbf{X},\mathcal{M}) = \frac{p(\mathbf{X}|\mathbf{W},\mathcal{M})p(\mathbf{W}|\mathcal{M})}{p(\mathbf{X}|\mathcal{M})} \tag{S1}$$

where

$$p(\mathbf{X}|\mathcal{M}) = \int p(\mathbf{X}|\mathbf{W},\mathcal{M})p(\mathbf{W}|\mathcal{M})\mathrm{d}\mathbf{W} \tag{S2}$$

In most of the cases, the integral at right hand side (RHS) of (S2) is intractable because it involves the integration in the whole weight space. Such an integral is also called as *evidence*, and it represents the pdf of the observed data **X** under the model $\mathcal{M}$. The variational approach allows us to approximate the integral in (S2), and the approximate form of the posterior pdf of the weights **W**, $p'(\mathbf{W})$, is introduced to allow a closed form solution to the posterior (left hand side, LHS, of S1). The idea behind this approach is the following. Let us consider the log-evidence $\ln(p(\mathbf{X}))$, where we drop the symbol $\mathcal{M}$ for brevity. Since the integral over the whole weight space of any given pdf, $p'(\mathbf{W})$, that depends only on the weights must be equal to the identity, and since the log-evidence does not depend on the weights **W**, the following equivalences hold:

$$\ln(p(\mathbf{X})) = \ln(p(\mathbf{X}))\int p'(\mathbf{W})\mathrm{d}\mathbf{W} = \int p'(\mathbf{W})\ln(p(\mathbf{X}))\mathrm{d}\mathbf{W} \tag{S3}$$

Using the standard formulas for joint pdfs: $p(\mathbf{X}, \mathbf{W}) = p(\mathbf{X} \mid \mathbf{W})\, p(\mathbf{W}) = p(\mathbf{W} \mid \mathbf{X})\, p(\mathbf{X})$, we can write:

$$\begin{aligned} \ln(p(\mathbf{X})) &= \int p'(\mathbf{W})\ln(p(\mathbf{X}))\mathrm{d}\mathbf{W} = \\ &= \int p'(\mathbf{W})\ln\left(\frac{p(\mathbf{X},\mathbf{W})}{p(\mathbf{W}|\mathbf{X})}\right)\mathrm{d}\mathbf{W} = \int p'(\mathbf{W})\ln\left(\frac{p(\mathbf{X},\mathbf{W})}{p(\mathbf{W}|\mathbf{X})}\frac{p'(\mathbf{W})}{p'(\mathbf{W})}\right)\mathrm{d}\mathbf{W} \end{aligned} \tag{S4}$$

Finally, rearranging equation (S4), we can write the following equality:

$$\ln(p(\mathbf{X})) = \int p'(\mathbf{W})\ln\left(\frac{p(\mathbf{X},\mathbf{W})}{p'(\mathbf{W})}\right)\mathrm{d}\mathbf{W} + \int p'(\mathbf{W})\ln\left(\frac{p'(\mathbf{W})}{p(\mathbf{W}|\mathbf{X})}\right) \tag{S5}$$

The first term of the RHS is called the Negative Free Energy of the data **X**, NFE[**X**], while the second term of the RHS is the Kullback-Leibler (KL-)divergence, KL[$p'(\mathbf{W}) \parallel p(\mathbf{W} \mid \mathbf{X})$], between the two pdfs $p'(\mathbf{W})$ and $p(\mathbf{W} \mid \mathbf{X})$, . The KL-divergence between two given pdfs is a strictly non-negative quantity that measures the difference among the pdfs under comparison, and it is equal to 0 iff the two pdfs are the same. In particular, the smallest the KL-divergence the more similar are the two pdfs. In the case under study, our goal is to find a pdf $p'(\mathbf{W})$ such that it approximates well the posterior pdf $p(\mathbf{W} \mid \mathbf{X})$. In other words, we want to minimize RHS's second term of equation (S5). In order to achieve this goal, since the log-evidence does not depend on the weights **W**, maximizing the NFE w.r.t. the $p'(\mathbf{W})$ will automatically minimize the KL-divergence. Let us now discuss further the structure of the NFE.

We can write the NFE as follows:

$$\begin{aligned} \mathrm{NFE}[\mathbf{X}] &= \int p'(\mathbf{W})\ln\left(\frac{p(\mathbf{X},\mathbf{W})}{p'(\mathbf{W})}\right) = \\ &= \int p'(\mathbf{W})\ln(p(\mathbf{X},\mathbf{W}))\mathrm{d}\mathbf{W} - \int p'(\mathbf{W})\ln p'(\mathbf{W})\mathrm{d}\mathbf{W} = \\ &= \langle \ln(p(\mathbf{X},\mathbf{W}))\rangle_{p'(\mathbf{W})} + H[\mathbf{W}] \end{aligned} \tag{S6}$$

where $<\cdot>_{p'(\mathbf{W})}$ is the expected value given the pdf $p'(\mathbf{W})$, and $H[\mathbf{W}]$ is the entropy of $p'(\mathbf{W})$. The most common restrictions for p' are (Ormerod and Wand, 2010):

a) $p'(\mathbf{W})$ factorizes into $\prod_{i=1}^{N} p'(w_i)$, for some partition $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ of **W**

b) p' is a member of a parametric family of density functions.

From equation (S6) it becomes clear the reason why it is necessary to choose a proper factorization for $p'(\mathbf{W})$ if we want to be able to maximize the NFE w.r.t. $p'(\mathbf{W})$. Indeed, such a maximiza-

tion is performed iteratively via an Expectation-Maximization procedure, and two different factorizations result in two different sets of learning rules.

For the particular case of the BSS problem, and for the implementation of a variational bayesian ICA, Choudrey (2002) used the ensemble of hidden variables and parameters given by $\mathbf{W} = \{\mathbf{\Lambda}, \mathbf{A}, \mathbf{S}, \mathbf{q}, \boldsymbol{\theta}\}$. He tested two different factorizations:

$$\text{i) } p'(\mathbf{W}) = p'(\mathbf{\Lambda})p'(\mathbf{A})p'(\mathbf{S})p'(\mathbf{q})p'(\boldsymbol{\theta}) \tag{S7}$$

$$\text{ii) } p'(\mathbf{W}) = p'(\mathbf{\Lambda})p'(\mathbf{A})p'(\mathbf{S} \mid \mathbf{q})p'(\mathbf{q})p'(\boldsymbol{\theta}) \tag{S8}$$

where the factorization ii) outperforms the first one (Choudrey, 2002). This is the reason why in our work we use only the second factorization. The meaning of the rvs can be understood taking a look to the original BSS problem, formulated by the equation (2) of the main text. We want to find those parameters (i.e., those weights $\mathbf{W}$) that can explain the data $\mathbf{X}$ under the framework of linear combination of independent source signals. The sources are described by the rvs $\mathbf{S}$, the mixing matrix is described by the rvs $\mathbf{A}$, and the noise is described by the rvs $\mathbf{\Lambda}$. Each of these rvs can be described by a given distribution or by other rvs.

Here we are going to present the so called learning rules we apply in order to maximize the NFE. In particular we specify the approximating posteriors in the RHS of (S8) used in order to calculate explicitly the NFE. MacKay (1995) has shown that there is no need to specify functional forms for the posteriors if conjugate forms for the densities are chosen. Families of conjugate functions are such that, when member functions are multiplied together, they give a function in the same family.

Noise is assumed to be Gaussian with zero mean, and the rvs $\mathbf{\Lambda}$ describe the precision (i.e., the inverse of the variance) associated to the noise. Such rvs are assumed to follow a Gamma distri-

bution, such that two hyper-parameters are necessary to describe each of them. The distribution of the mixing matrix coefficients is a Gaussian, where each element of the matrix **A** is described by a mean and a precision. Finally, in order to allow the sources to adapt and mimic different distributions, the rvs **S** are described using a mix of Gaussian distributions for each source $\mathbf{s}_i$ with $i = 1, \ldots, L$. This means that for each source $\mathbf{s}_i$ there is a set of rvs $\boldsymbol{\theta}_i$ such that it explains the set of Gaussian pdfs that contributes to the final realization of $\mathbf{s}_i$. Supposing that m_i Gaussian pdfs are used to describe the i-th source, then we need a mean and a precision for each of the m_i Gaussians as well as a rv π_i expressing the probability that a given Gaussian is selected to contribute to the source. Finally, the rv **q** is an indicator variable, and for the i-th source there is the rv q_i that may vary between 1 and m_i. The mean of a Gaussian μ_{i,q_i} follows itself a normal distribution, while a precision of a Gaussian β_{i,q_i} follows a Gamma distribution. The mixture proportion π_i is described by a Dirichlet distribution.

The choice of these particular families of distribution is arbitrary, but not accidental. Indeed, the Gamma distribution ($\mathscr{G}(\cdot)$) is the conjugate prior for the precision of the normal distribution ($\mathscr{N}(\cdot)$) with known mean, and the Dirichlet distribution ($\mathscr{D}(\cdot)$) is the conjugate prior of the categorical distribution. All the previous assumptions can be summarized by Figure S1, which represents the direct graph to solve the BSS problem via an ICA. The formulas associated to all these assumptions on the prior distributions are the following:

$$p(\boldsymbol{\Lambda}) = \prod_{j=1}^{M} \mathscr{G}(\Lambda_j; b_{\Lambda_j}, c_{\Lambda_j}) \tag{S9}$$

$$p(\mathbf{A}) = \prod_{i=1}^{L} \prod_{j=1}^{M} \mathscr{N}(A_{ji}|0, \alpha_{ji}) \tag{S10}$$

$$p(\mathbf{S}|\mathbf{q}^{\mathbf{t}}, \boldsymbol{\theta}) = \prod_{t=1}^{T} \prod_{i=1}^{L} \mathscr{N}(s_i^t; \mu_{i,q_i}, \beta_{i,q_i}) \tag{S11}$$

$$p(\mathbf{q}|\boldsymbol{\pi}) = \prod_{t=1}^{T}\prod_{i=1}^{L}\pi_{i,q_i} \tag{S12}$$

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}) \tag{S13}$$

$$p(\boldsymbol{\pi}) = \prod_{i=1}^{L}\mathscr{D}\,(\boldsymbol{\pi_i};\lambda_{i0}) \tag{S14}$$

$$p(\boldsymbol{\mu}) = \prod_{i=1}^{L}\prod_{q_i=1}^{m_i}\mathscr{N}\,(\mu_{i,q_i};m_{i0},\tau_{i0}) \tag{S15}$$

$$p(\boldsymbol{\beta}) = \prod_{i=1}^{L}\prod_{q_i=1}^{m_i}\mathscr{G}\,(\beta_{i,q_i};b_{i0},c_{i0}) \tag{S16}$$

The derivation of the posteriors can be found in the Appendix B and C of Choudrey (2002). Here we just report his results, in order to describe the modifications we have introduced for applications to cases with missing data, following Chan et al. (2003). The priors from (S9) to (S16) become the posteriors given by:

$$p'(\boldsymbol{\Lambda}) = \prod_{j=1}^{M}\mathscr{G}\,(\Lambda_j;\hat{b}_{\Lambda_j},\hat{c}_{\Lambda_j}) \tag{S17}$$

$$p'(\mathbf{A}) = \prod_{i=1}^{L}\prod_{j=1}^{M}\mathscr{N}\,(A_{ji}|\widehat{m}_{A_{ji}},\widehat{\alpha}_{ji}) \tag{S18}$$

$$p'(\mathbf{S}|\mathbf{q^t},\boldsymbol{\theta}) = \prod_{t=1}^{T}\prod_{i=1}^{L}\mathscr{N}\,(s_i^t;\hat{\mu}_{i,q_i}^{(t)},\hat{\beta}_{i,q_i}^{(t)}) \tag{S19}$$

$$p'(\mathbf{q}) = \prod_{t=1}^{T}\prod_{i=1}^{L}\hat{\gamma}_{i,q_i}^{(t)} \tag{S20}$$

$$p'(\boldsymbol{\theta}) = p'(\boldsymbol{\pi})p'(\boldsymbol{\mu})p'(\boldsymbol{\beta}) \tag{S21}$$

$$p'(\boldsymbol{\pi}) = \prod_{i=1}^{L} \mathscr{D}\,(\boldsymbol{\pi_i}; \hat{\lambda}_{i,q:m_i}) \tag{S22}$$

$$p'(\boldsymbol{\mu}) = \prod_{i=1}^{L} \prod_{q_i=1}^{m_i} \mathscr{N}\,(\mu_{i,q_i}; \widehat{m}_{i,q_i}, \hat{\tau}_{i,q_i}) \tag{S23}$$

$$p'(\boldsymbol{\beta}) = \prod_{i=1}^{L} \prod_{q_i=1}^{m_i} \mathscr{G}\,(\beta_{i,q_i}; \hat{b}_{i,q_i}, \hat{c}_{i,q_i}) \tag{S24}$$

From these approximating pdfs, it is possible to maximize the NFE w.r.t. the $p'(\mathbf{w}_i)$. The parameters are estimated using an Expectation-Maximization algorithm, obtaining the values indicated with the hat ˆ at each iteration. The missing data are taken into account using a data mask o_j^t, with $j = 1, \ldots, M$ and $t = 1, \ldots, T$, that is equal to 0 if the data is missing and 1 if the data is recorded. The learning rules for the different pdf hyper-parameters are the following:

$p'(\boldsymbol{\Lambda})$

$$\hat{b}_{\Lambda_j} = \left[\frac{1}{b_{\Lambda_j}} + \frac{1}{2}\sum_{t=1}^{T} o_j^t \langle (x_j^t - \hat{x}_j^t)^2 \rangle\right]^{-1} \tag{S25}$$

$$\hat{c}_{\Lambda_j} = c_{\Lambda_j} + \frac{1}{2}\sum_{t=1}^{T} o_j^t \tag{S26}$$

$p'(\mathbf{A})$

$$\widehat{m}_{A_{ji}} = \frac{\langle \Lambda_j \rangle}{\widehat{\alpha}_{ji}} \sum_{t=1}^{T} \langle s_i^t \rangle\, o_j^t (x_j^t - \langle \hat{x}_{j,k\neq i}^t \rangle) \tag{S27}$$

$$\widehat{\alpha}_{ji} = \alpha_{ji} + \langle \Lambda_j \rangle \sum_{t=1}^{T} o_j^t \langle {s_i^t}^2 \rangle \tag{S28}$$

$p'(\mathbf{S} \mid \mathbf{q})$

$$\hat{\mu}_{i,q_i}^t = \frac{1}{\hat{\beta}_{i,q_i}^t}\left[\langle\beta_{i,q_i}\rangle\langle\mu_{i,q_i}\rangle + \sum_{j=1}^{M}\langle\Lambda_j\rangle\langle A_{ji}\rangle o_j^t\left(x_j^t - \langle\hat{x}_{j,k\neq i}^t\rangle\right)\right] \tag{S29}$$

$$\hat{\beta}_{i,q_i}^t = \langle\beta_{i,q_i}\rangle + \sum_{j=1}^{M} o_j^t\langle\Lambda_j\rangle\langle A_{ji}^2\rangle \tag{S30}$$

$p'(\mathbf{q})$

$$\hat{\gamma}_{i,q_i}^t = \frac{\gamma_{i,q_i}^t}{\sum_{q_i'}\gamma_{i,q_i}^t} \tag{S31}$$

where $\gamma_{i,q_i}^t = \tilde{\pi}_{i,q_i}\tilde{p}_{i,q_i}$ and

$$\tilde{\pi}_{i,q_i} = \exp\left[\Psi(\hat{\lambda}_{i,q_i}) - \Psi\left(\sum_{q'_i}\hat{\lambda}_{i,q'_i}\right)\right] \tag{S32}$$

$$\tilde{p}_{i,q_i} = \left(\frac{\tilde{\beta}_{i,q_i}}{\hat{\beta}_{i,q_i}^t}\right)^{\frac{1}{2}}\exp\left[\frac{1}{2}\left(\hat{\beta}_{i,q_i}^t\hat{\mu}_{i,q_i}^{t\ 2} - \langle\beta_{i,q_i}\rangle\langle\mu_{i,q_i}^2\rangle\right)\right] \tag{S33}$$

$$\tilde{\beta}_{i,q_i} = \hat{b}_{i,q_i}\exp\left[\Psi(\hat{c}_{i,q_i})\right] \tag{S34}$$

$p'(\boldsymbol{\pi})$

$$\hat{\lambda}_{i,q_i} = \lambda_{i0} + \sum_{t=1}^{T}\hat{\gamma}_{i,q_i}^t \tag{S35}$$

$p'(\boldsymbol{\mu})$

$$\hat{m}_{i,q_i} = \frac{1}{\hat{\tau}_{i,q_i}}\left(\tau_{i0}m_{i0} + \langle\beta_{i,q_i}\rangle\sum_{t=1}^{T}\hat{\gamma}_{i,q_i}^t\langle s_i^t|q_i^t\rangle\right) \tag{S36}$$

$$\hat{\tau}_{i,q_i} = \tau_{i0} + \langle \beta_{i,q_i} \rangle \sum_{t=1}^{T} \hat{\gamma}_{i,q_i}^{t} \tag{S37}$$

$p'(\boldsymbol{\beta})$

$$\hat{b}_{i,q_i} = [\frac{1}{b_{i0}} + \frac{1}{2}\tilde{\sigma}_{i,q_i}]^{-1} \tag{S38}$$

$$\hat{c}_{i,q_i} = c_{i0} + \frac{1}{2}\sum_{t=1}^{T} \hat{\gamma}_{i,q_i}^{t} \tag{S39}$$

where $\tilde{\sigma}_{i,q_i} = \sum_{t=1}^{T} \hat{\gamma}_{i,q_i}^{t} \left(\langle {s_i^t}^2 | q_i^t \rangle - 2\langle \mu_{i,q_i} \rangle \langle s_i^t | q_i^t \rangle + \langle \mu_{i,q_i}^2 \rangle \right)$.

We specify also the learning rules relative to the Automatic Relevance Determination technique (see Section 4 of the main text). Indeed, the precision of each column of the mixing matrix is treated as a rv, and it follows its own distribution. In particular, it obeys a Gamma distribution:

$$p'(\boldsymbol{\alpha}) = \prod_{i=1}^{L} \mathrm{G}(\alpha_i; \hat{b}_{\alpha_i}, \hat{c}_{\alpha_i}) \tag{S40}$$

and the updating equations for the parameters $\hat{b}_{\alpha_i}$ and $\hat{c}_{\alpha_i}$ are:

$$\hat{b}_{\alpha_i} = [\frac{1}{b_{\alpha_i}} + \frac{1}{2}\sum_{j=1}^{M} \langle A_{ji}^2 \rangle]^{-1} \tag{S41}$$

$$\hat{c}_{\alpha_i} = c_{\alpha_i} + \frac{M}{2} \tag{S42}$$

In all case studies we have used the following starting hyper-parameter values: $m_{i0} = 0$, $\tau_{i0} = 1$, $b_{i0} = 10^3$, $c_{i0} = 10^{-3}$, $\lambda_{i0} \in [0.01, 0.1]T$, $b_{\alpha_i} = 10^3$, $c_{\alpha_i} = 10^{-3}$ for $i = 1, \ldots, L$; and $b_{\Lambda_j} = 10^3$, $c_{\Lambda_j} = 10^{-3}$ for $j = 1, \ldots, M$.

S2 – Simulated sources

S2.1 – Seismic cycle

In order to reproduce the seismic cycle, we use three different source signals: 1) a linear function; 2) a Heaviside function; 3) a logarithmic function, representing the inter-, co-, and post-seismic stages, respectively. The reason behind the choice of these functions can be found in Section 5.1 of the main text. The mathematical expressions used are the following:

$$s^{lin}(x,y,z,t) = q^{lin}(x,y,z) + m^{lin}(x,y,z)t \tag{S43}$$

$$s^{co}(x,y,z,t,t^{co}) = A^{co}(x,y,z)H(t-t^{co}) \tag{S44}$$

$$s^{post}(x,y,z,t,t^{co},\tau) = A^{post}(x,y,z)\ln(1+\frac{t-t^{co}}{\tau}) \tag{S45}$$

where t is time, and (x, y, z) is a point in the space.

We discretize the fault plane into different patches of smaller area. For each patch located in (x, y, z) we specify the following six parameters in (S43-S45)

1) q^{lin}: arbitrary value of stationary aseismic slip (creep) at time T_{start}

2) $m^{lin} = \dot{s}^{lin}$: creep rate

3) A^{co}: co-seismic slip

4) t^{co}: epoch of the earthquake

5) A^{post}: post-seismic amplitude

6) τ: post-seismic decay time.

Varying these 6 parameters we can vary the Signal-to-Signal Ratio (SSR) among the different sources, defined, in analogy with the Signal-to-Noise Ratio (SNR), as the ratio between the power of one signal and the power of a second signal. In particular, we maintain the same value for the fol-

lowing three parameters: q^{lin} = 0 mm, t^{co} = 1096 d (= 3 yr) and τ = 1 d. We limit the cases under study allowing the remaining three parameters to assume the values shown in Table S1.

The co-seismic and post-seismic amplitudes are intended to vary in order to generate events with different energies. Assuming that the fault under study is embedded in a homogeneous and elastic half-space, then the seismic moment associated to an earthquake (or the equivalent seismic moment associated to afterslip) can be calculated using the formula:

$$M_0 = \mu A \delta \tag{S46}$$

where μ is the rigidity modulus, and A is the area that slips an amount δ during the co-seismic period (or during the post-seismic period). In all our simulations, we keep constant the rigidity modulus μ = 30 GPa (a typical value for the crust, e.g. Kanamori and Brodsky, 2004). For the purposes of this work, it is sufficient to use uniform slip distributions in a set of contiguous patches within a certain depth interval, so the co-seismic and post-seismic amplitudes take only the values shown in Table S1. Since the moment magnitude is related to the seismic moment (in the International System of units) by the Hanks and Kanamori (1979) formula

$$M_w = \frac{2}{3}\log(M_0) - 6 \tag{S47}$$

then we vary the moment magnitude of the generated events by changing M_0. In turns we vary M_0 by changing the extension of the slipping portion of the fault and then its area A. Figure 2 of the main text shows the three different fault models proposed. We use a fixed planar fault geometry, described by the 7 parameters of Table S2. The tectonic regime is set to simulate a thrust fault (rake -90°). The associated M_w^{co} are 6.85, 6.29, and 5.94, respectively, and the corresponding equivalent M_w^{post} after 2 years are 6.57, 6.09, and 5.80.

S2.2 – Volcanic source

The time dependence of the magma chamber's volume associated to the volcanic source is modeled using the following equation:

$$V(t; a_{infl}, t_{infl}, a_{defl}, t_{defl}) \propto \frac{\arctan[a_{infl}(t - t_{infl})]}{max\{\arctan[a_{infl}(t - t_{infl})]\}} - \frac{\arctan[a_{defl}(t - t_{defl})]}{max\{\arctan[a_{defl}(t - t_{defl})]\}} \qquad \text{(S48)}$$

The values of the 4 parameters a_{infl}, t_{infl}, a_{defl}, t_{defl} are 0.01 1/yr, 4 yr, 0.05 1/yr, and 4.5 yr, respectively.

S3 - Credibility intervals for the post-seismic decay constant

Let us suppose that we want to explain a data vector $\mathbf{d} \in \mathbb{R}^D$ using a parameter vector $\mathbf{m} \in \mathbb{R}^M$ and the relationship (model) $g : \mathbb{R}^M \rightarrow \mathbb{R}^D$ such that $\mathbf{d} = g(\mathbf{m})$. The misfit function $S(\mathbf{m}, \mathbf{d}) = \mathbf{d} - g(\mathbf{m})$ depends on both the data and the parameter model vectors, **d** and **m**. For brevity we will consider the observed data as a known vector $\mathbf{d} = \mathbf{d}_{obs}$, and the misfit function depending only on the choice of the parameter vector. In general, g can be a non-linear function, and it is not guaranteed that S is convex everywhere in the model space. In other words, it is not guaranteed that only one minimum exists for the misfit function.

A bayesian approach to the fitting problem consists in assigning to the data **d**, the parameters **m**, and the model function g some *a priori* probability density function (pdf), and then solve for the inverse problem to find the *a posteriori* pdf of the parameters. In the particular fitting problem treated in this work, the observed data vector $\mathbf{d}_{obs}$ corresponds to the IC related to the post-seismic decay, the model function g is given by a slightly modified version of equation (S45):

$$g(\mathbf{m}, t, t^{co}) = m_1 + m_2 \ln(1 + \frac{t - t^{co}}{m_3}) \qquad \text{for } t > t^{co} \qquad \text{(S49)}$$

where $\mathbf{m}=(m_1, m_2, m_3)$, m_1 bias, m_2 amplitude, and m_3 decay time. We consider only times $t > t^{co}$, i.e. epochs after the occurrence of the mainshock. The time is a known parameter, so it can be neglected in the count of the model space dimensionality. We assume that modelization uncertainties are neg-

ligible compared to observational uncertainties. Since the data and model spaces are linear, the solution to the inverse problem can be written as (equation 1.93 of Tarantola, 2005):

$$\sigma_M(\mathbf{m}) = \frac{1}{\nu}\rho_M(\mathbf{m})\rho_D(g(\mathbf{m})) \tag{S50}$$

where σ_M is the *a posteriori* probability density of the parameters, ρ_M is the *prior* probability density in the model space, ρ_D is the probability density describing the result of the measurement, and $\nu = \int_{\mathbb{R}^M} \rho_M(\mathbf{m})\rho_D(g(\mathbf{m}))\mathrm{d}\mathbf{m}$ is a normalization constant. We will identify $\rho_D(g(\mathbf{m}))$ with the likelihood function $L(\mathbf{m})$, which gives a measure of how good a model $\mathbf{m}$ is in explaining the data.

If no *a priori* information is available about the parameters, $\rho_M(\mathbf{m})$ can be replaced by its homogeneous limit $\mu_M(\mathbf{m})$. We consider a uniform $\mu_M(\mathbf{m})=k$, and we can rewrite equation (S45) as:

$$\sigma_M(\mathbf{m}) = \frac{k}{\nu}L(\mathbf{m}) \tag{S51}$$

Knowing the *a posteriori* pdf of the parameters $\mathbf{m}$, we can compute the x% credibility volume V_x in the M-dimensional space, that is determined as the volume V_x such that the probability P to find a parameter vector in it is x%, i.e. $V_x = V \subseteq \mathbb{R}^M$ such that $P(\mathbf{m} \in V) = x\%$. From the definition of probability we can write:

$$x\% = P(\mathbf{m} \in V_x) = \int_{V_x} \sigma_M(\mathbf{m})\mathrm{d}\mathbf{m} = \frac{\int_{V_x} L(\mathbf{m})\mathrm{d}\mathbf{m}}{\int_{\mathbb{R}^M} L(\mathbf{m})\mathrm{d}\mathbf{m}} \tag{S52}$$

Obviously, if $V_x = \mathbb{R}^M$, then the probability to find $\mathbf{m}$ in V_x is 100%. Since the problem is not linear, it is not possible to use a direct formula to solve it and we have to sample the model space. Fortunately, for all the cases treated in this work the model space dimensionality is low (M = 3), and we can sample the posterior distribution using a grid search approach. It remains to specify the a priori pdf for the data, $\rho_D(g(\mathbf{m}))$ or $L(\mathbf{m})$. This choice is case dependent. In the cases treated here, the data consist in the temporal sources obtained from an ICA. A great advantage of the vbICA

technique is that it finds an approximation of the sources using a mix of Gaussian distributions, and thus it is possible to calculate the moments of such a distribution. In particular, for each point of a given IC, i.e. for each element of the temporal sources, we can compute (at least) the variance. We assume that all the T elements that compose an IC are independent and identically distributed (iid), in particular they follow a normal distribution $\rho_D(g(\mathbf{m}))$. Practically, we are not assuming any temporal dependency between the value of the IC at time t and the value at time $t + \delta t$. An attempt to consider this dependency has been done by Choudrey (2002). He developed a Dynamic ICA, based on Hidden Markov Models (HMMvbICA). It might be a good idea for the future to follow the same approach also for the analysis of geodetic data.

We explore the ranges [-0.03,0], [ε,0.04], and [ε,50] for the three parameters m_1, m_2, and m_3, where $\varepsilon = 2.2 \times 10^{-16}$. These ranges are chosen around the values found using an unconstrained non-linear minimization of the sum of squared residuals, and ensure us to explore all the regions of the model space where the likelihood function is significantly greater than 0. We say that the likelihood is significantly greater than 0 if even considering a m_3 range 100 times wider, the sum of all the new added points investigated would not contribute more than 1 /10 of the contribution to the volume given by the point of maximum likelihood. If a relevant portion of the model space having a likelihood function significantly different from 0 is not taken into account, than the uncertainty on the time decay parameter is too high, and such a parameter is not resolved. The grid step adopted for the three parameters is 0.005, 0.0001, and 0.5 respectively. The actual m_3 value is 1 day (see Section 5.1 of the main text), and the credible intervals at 68.27% and 99.99% for the low tectonic rate case are listed in Tables S7 and S8. For 6 over 18 cases the m_3 parameter is not resolved, and all the values spanned are acceptable or the post-seismic source has not been identified (NI). These 6 cases correspond to the N2 geometry scenarios for the intermediate and small post-seismic source intensities. This proves that it is necessary to have a good quality network if we want to study crustal displacements of the order of few mm with multivariate statistical techniques such as vbICA. In more than one case we recover more than one credible interval. This is the direct consequence of

the non linearity of the forward model. Indeed, the likelihood function has few local maxima, and we can not exclude the possibility that the proper model does not belong to the global maximum. Moreover, as every inverse problem, the solution is not unique, and to give only the maximum likelihood model could be misleading. In all cases where the number of ICs used is 2, it is probably more correct to use a forward model (i.e., equation S49) that takes into account also the linear signal. Nevertheless, the number of parameters used in equation S49 seems to be already enough to correctly guess the decay constant at a 99.99% level.

Figure captions

Figure S1: Same as figure 5.1 of Choudrey (2002). Bayesian Independent Component Analysis as a Graphical Model. Circles represent random variables and rectangles represent hyper-parameters. The meaning of each symbol is indicated in the main text of the supplementary material. The mixing matrix parameters are all summarized in one symbol for brevity.

Figure S2: Black dots: temporal evolution of the recovered PCs (left) and ICs (right) in the case of 0%, 5%, and 25% of missing data. The gray shadow in the ICs corresponds to the associated uncertainty related to the ICs, calculated as the square root of the variance. This estimation is enabled by the knowledge of the approximated pdf of each IC via a mix of Gaussians. The red lines correspond to the post-seismic and seasonal actual sources. The scenario here represented corresponds to the one of the network geometry N1, a big post-seismic source $M_w^{post} = 6.57$, and a low tectonic rate $\dot{s}^{lin} = 2$ mm/yr.

Figure S3: As Figure S2, but with N2 – $M_w^{post} = 6.57$ – $\dot{s}^{lin} = 2$mm/yr.

Figure S4: As Figure S2, but with N1 – $M_w^{post} = 6.09$ – $\dot{s}^{lin} = 2$mm/yr.

Figure S5: As Figure S2, but with N2 – $M_w^{post} = 6.09$ – $\dot{s}^{lin} = 2$mm/yr.

Figure S6: As Figure S2, but with N1 – $M_w^{post} = 5.80$ – $\dot{s}^{lin} = 2$mm/yr.

Figure S7: As Figure S2, but with N2 – $M_w^{post} = 5.80$ – $\dot{s}^{lin} = 2$mm/yr.

Tables

Parameter	Values			Unit of measurement
m^{lin}	2	12	60	mm/yr
A^{co}	400	400	200	mm
A^{post}	30.3	15.2	9.1	mm

Table S1: Source parameters case study

$x_{top\ centre}$	0	km
$y_{top\ centre}$	0	km
Length	46	km
z_{top}	-2	km
z_{bottom}	-26	km
Strike	45	°
Dip	40	°

Table S2: Fault plane geometry parameters.

		$\dot{s}^{lin} = 2$ mm/yr			$\dot{s}^{lin} = 12$ mm/yr			$\dot{s}^{lin} = 60$ mm/yr		
		MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%
M_w^{co} 6.85	N1	0.26	0.26	0.24	9.45	9.50	8.58	236.32	237.54	214.40
	N2	0.22	0.21	0.19	7.97	7.73	6.84	199.17	193.29	170.91
M_w^{co} 6.29	N1	0.26	0.26	0.24	9.45	9.50	8.58	236.32	237.54	214.40
	N2	0.22	0.21	0.19	7.97	7.73	6.84	199.17	193.29	170.91
M_w^{co} 5.94	N1	0.26	0.26	0.24	9.45	9.50	8.58	236.32	237.54	214.40
	N2	0.22	0.21	0.19	7.47	7.73	6.84	199.17	193.29	170.91

Table S3: Linear SNR values.

		$\dot{s}^{lin} = 2$ mm/yr			$\dot{s}^{lin} = 12$ mm/yr			$\dot{s}^{lin} = 60$ mm/yr		
		MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%
M_w^{co} 6.85	N1	339.58	324.46	328.63	339.58	324.46	328.63	339.58	324.46	328.63
	N2	14.97	14.20	13.16	14.97	14.20	13.16	14.97	14.20	13.16
M_w^{co} 6.29	N1	12.34	12.08	11.54	12.34	12.08	11.54	12.34	12.08	11.54
	N2	0.85	0.83	0.70	0.85	0.83	0.70	0.85	0.83	0.70
M_w^{co} 5.94	N1	1.30	1.28	1.22	1.30	1.28	1.22	1.30	1.28	1.22
	N2	0.08	0.08	0.06	0.08	0.08	0.06	0.08	0.08	0.06

Table S4: Co-seismic SNR values

		$\dot{s}^{lin} = 2$ mm/yr			$\dot{s}^{lin} = 12$ mm/yr			$\dot{s}^{lin} = 60$ mm/yr		
		MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%
M_w^{co} 6.85	N1	64.30	63.35	63.04	64.30	63.35	63.04	64.30	63.35	63.04
	N2	4.43	4.19	3.37	4.43	4.19	3.37	4.43	4.19	3.37
M_w^{co} 6.29	N1	2.22	2.11	1.84	2.22	2.11	1.84	2.22	2.11	1.84
	N2	0.10	0.09	0.07	0.10	0.09	0.07	0.10	0.09	0.07
M_w^{co} 5.94	N1	0.31	0.29	0.26	0.31	0.29	0.26	0.31	0.29	0.26
	N2	0.013	0.012	0.009	0.013	0.012	0.009	0.013	0.012	0.009

Table S5: Post-seismic SNR values

		$\dot{s}^{lin} = 2$ mm/yr			$\dot{s}^{lin} = 12$ mm/yr			$\dot{s}^{lin} = 60$ mm/yr		
		MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%	MD 0%	MD 5%	MD 25%
M_w^{co} 6.85	N1	0.98	0.98	0.97	0.98	0.98	0.97	0.98	0.98	0.97
	N2	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
M_w^{co} 6.29	N1	0.98	0.98	0.97	0.98	0.98	0.97	0.98	0.98	0.97
	N2	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
M_w^{co} 5.94	N1	0.98	0.98	0.97	0.98	0.98	0.97	0.98	0.98	0.97
	N2	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11

Table S6: Seasonal SNR values.

m_3 (days)		$\dot{s}^{lin}$ = 2 mm/yr		
		MD 0%	MD 5%	MD 25%
M_w^{co} = 6.85	N1	[1.0]	[1.0]	[1.0]
	N2	[2.0, 2.5]	[1.5]	[0.5]
M_w^{co} = 6.29	N1	[1.5], [4.0]	[1.0]	[1.0]
	N2	[15.5, 37.5]	[5.0],[6.0,9.5], [10.5],[38.0,50.0]	[NI]
M_w^{co} = 5.94	N1	[11.5, 15.5]	[10.5, 15.5]	[0.5,1.0]
	N2	[NI]	[NI]	[ε, 50.0]

Table S7: m_3 ranges in days: 68.27% credible intervals.

m_3 (days)		$\dot{s}^{lin}$ = 2 mm/yr		
		MD 0%	MD 5%	MD 25%
M_w^{co} = 6.85	N1	[1.0]	[1.0]	[1.0]
	N2	[1.0],[2.0,2.5], [4.5,5.5]	[1.5]	[0.5, 1.0]
M_w^{co} = 6.29	N1	[1.5,2.0],[3.0,5.0]	[0.5,1.0],[2.0,3.0]	[0.5, 1.0], [2.0]
	N2	[0.5, 50]	[0.5,50.0]	[NI]
M_w^{co} = 5.94	N1	[0.5,1.5],[2.5,5.5], [8.0,21.5]	[0.5,5.5], [7.0,22.5]	[0.5,1.5]
	N2	[NI]	[NI]	[ε, 50.0]

Table S8: m_3 ranges in days: 99.99% credible intervals.

Figures

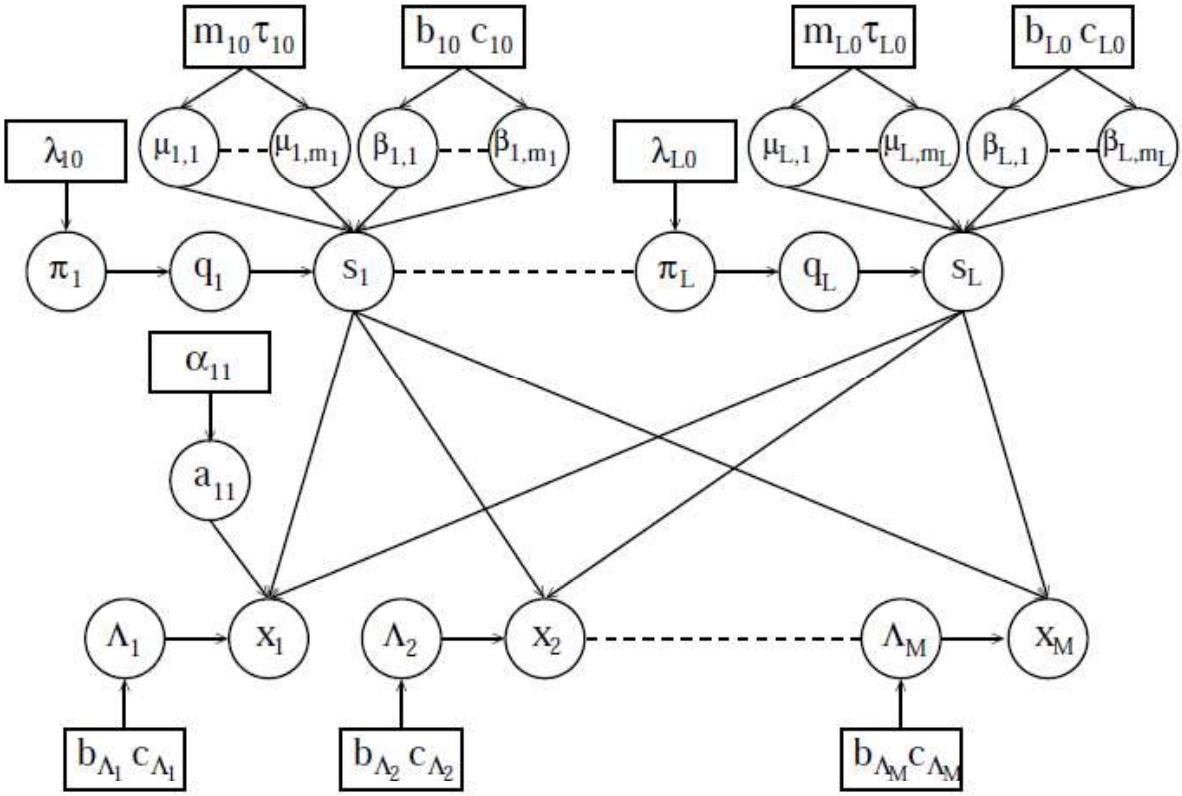
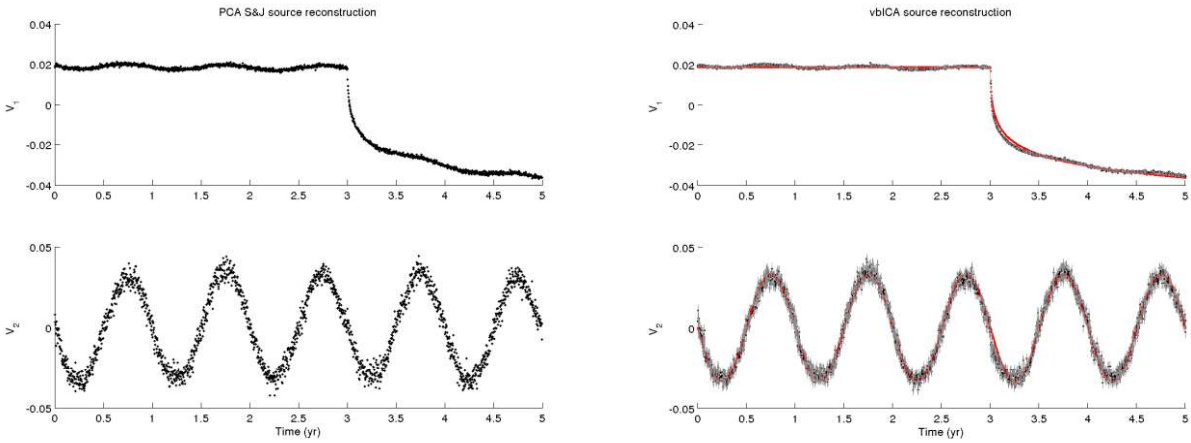
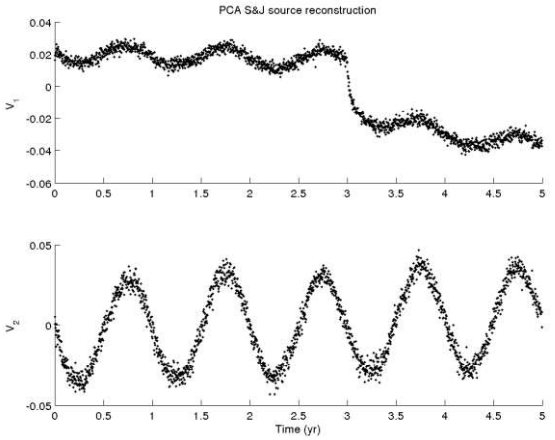


Figure S1

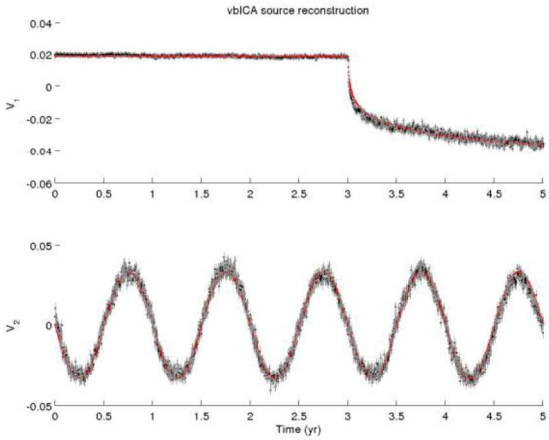


a) PCA – MD 0%

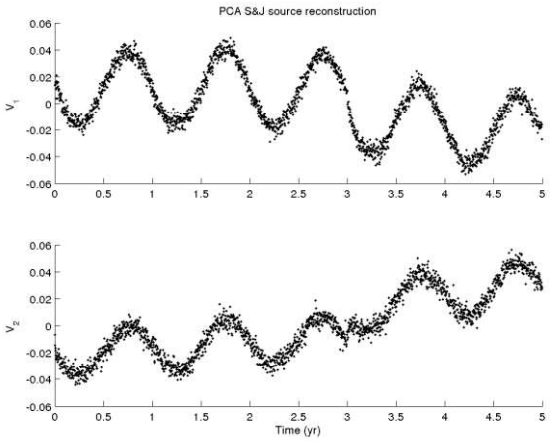
b) vbICA – MD 0%



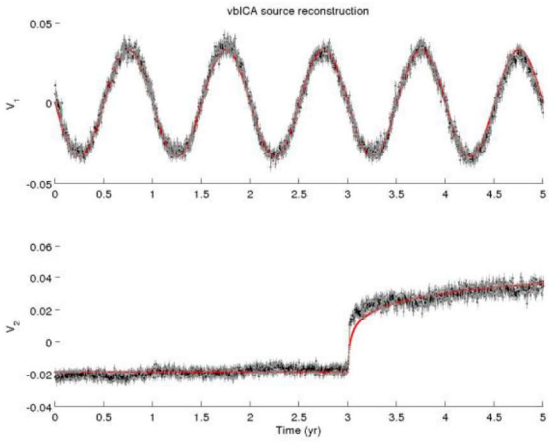
c) PCA – MD 5%



d) vbICA – MD 5%

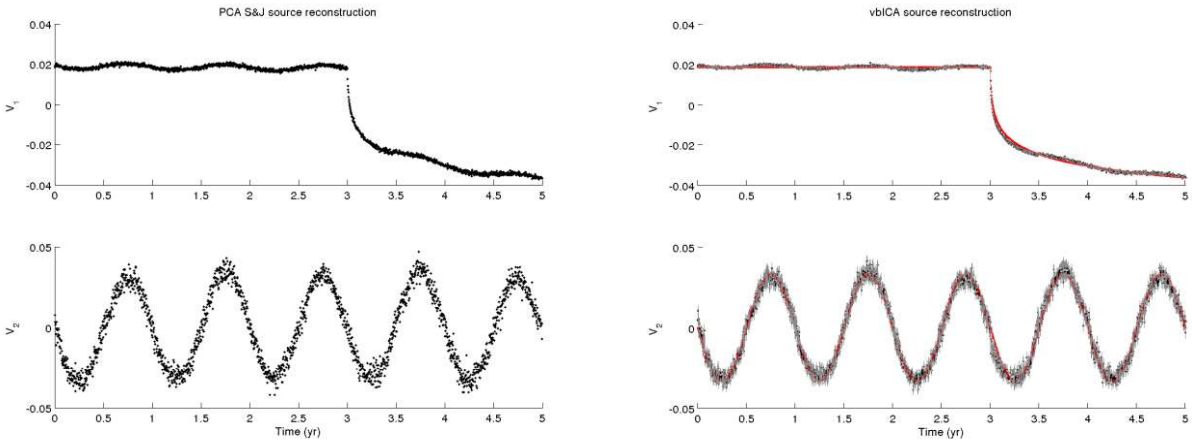


e) PCA – MD 25%



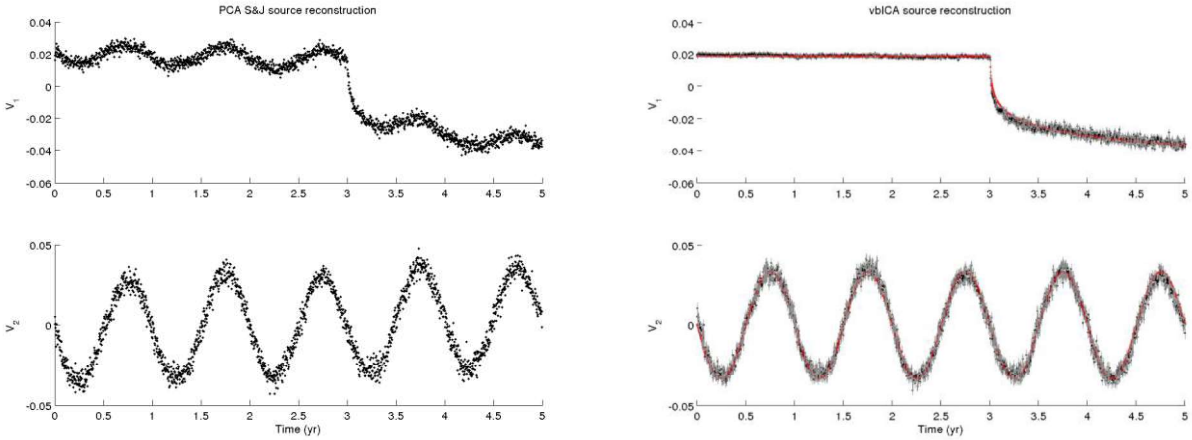
f) vbICA – MD 25%

Figure S2



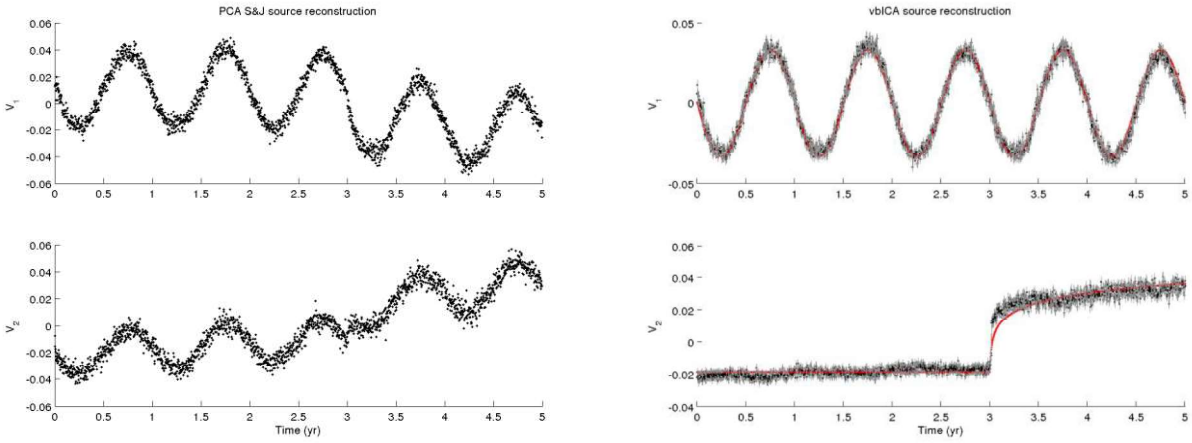
a) PCA – MD 0%

b) vbICA – MD 0%



c) PCA – MD 5%

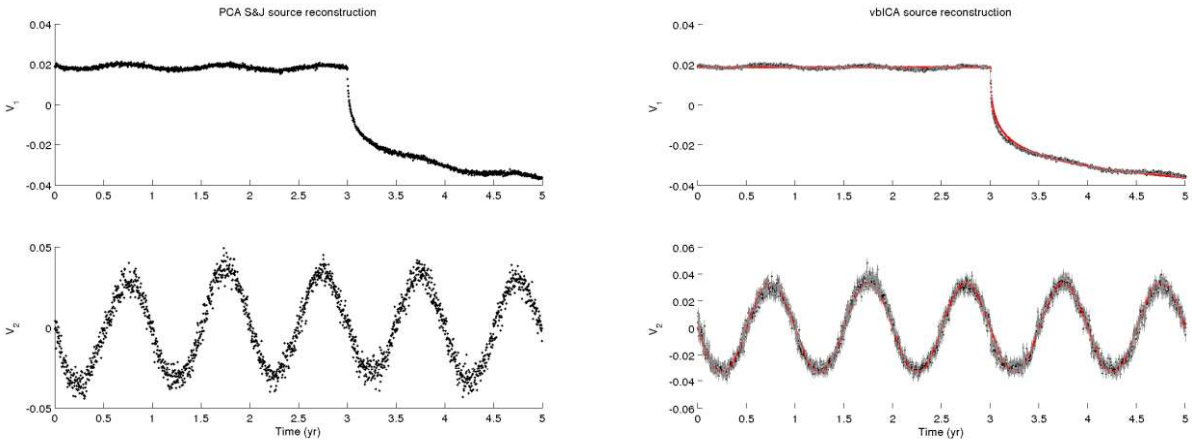
d) vbICA – MD 5%



e) PCA – MD 25%

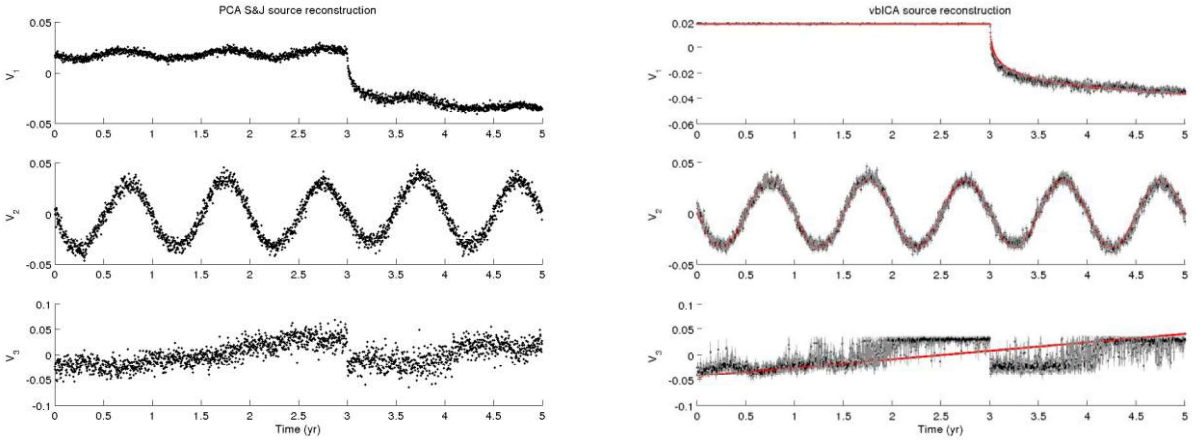
f) vbICA – MD 25%

Figure S3



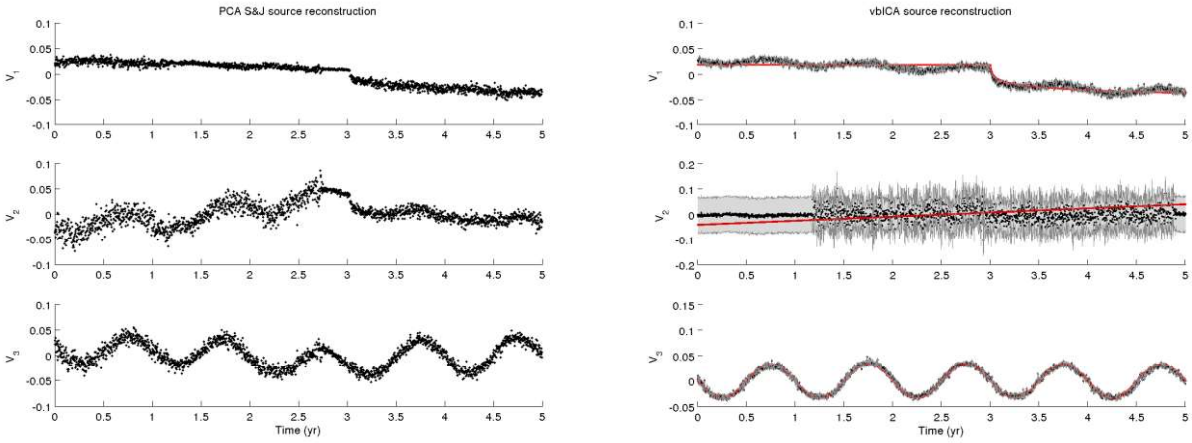
a) PCA – MD 0%

b) vbICA – MD 0%



c) PCA – MD 5%

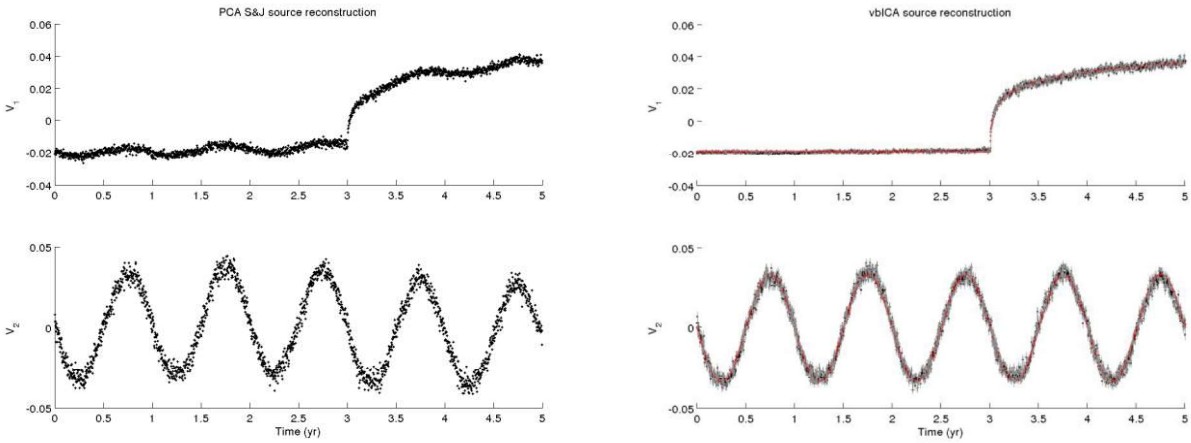
d) vbICA – MD 5%



e) PCA – MD 25%

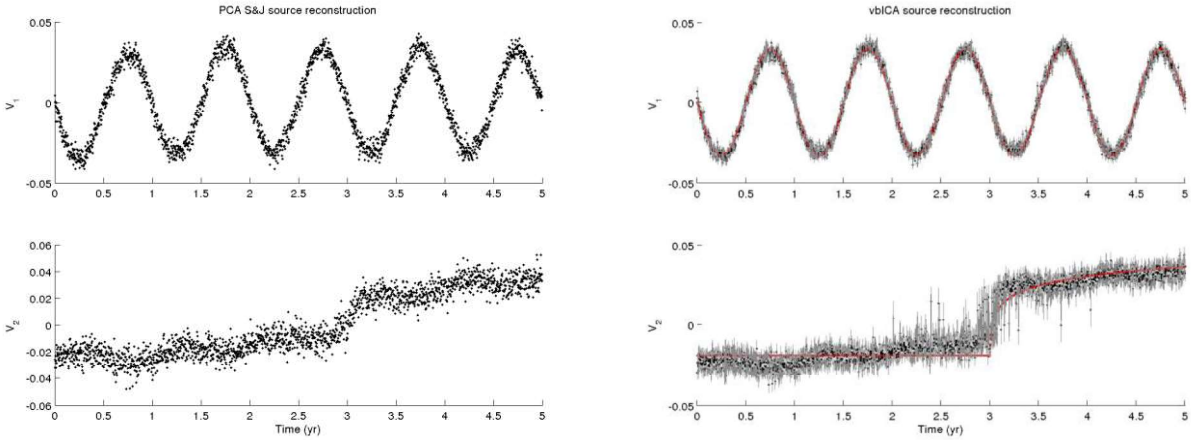
f) vbICA – MD 25%

Figure S4



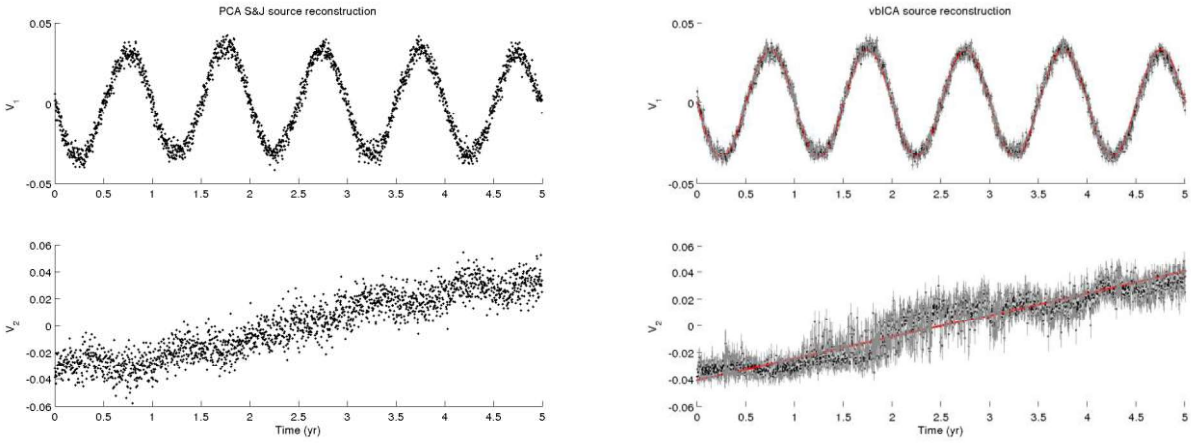
a) PCA – MD 0%

b) vbICA – MD 0%



c) PCA – MD 5%

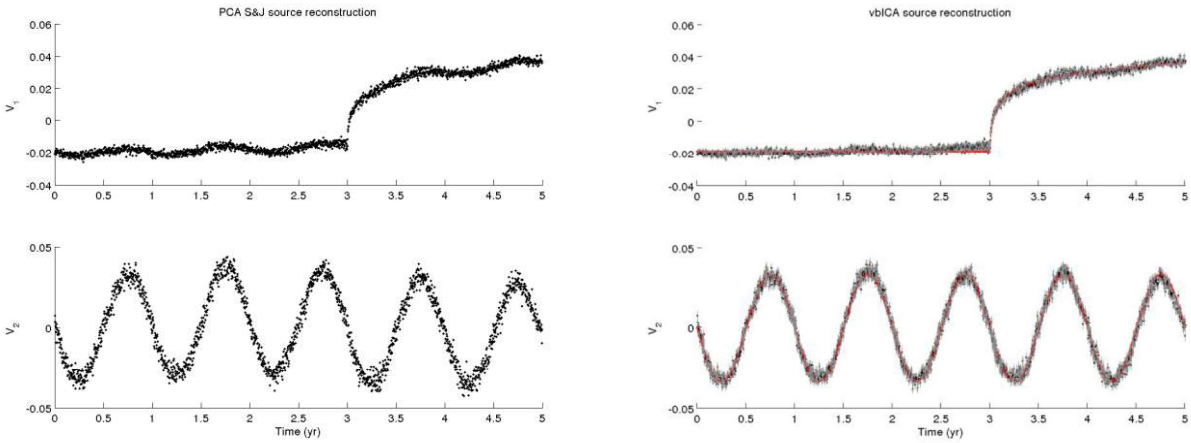
d) vbICA – MD 5%



e) PCA – MD 25%

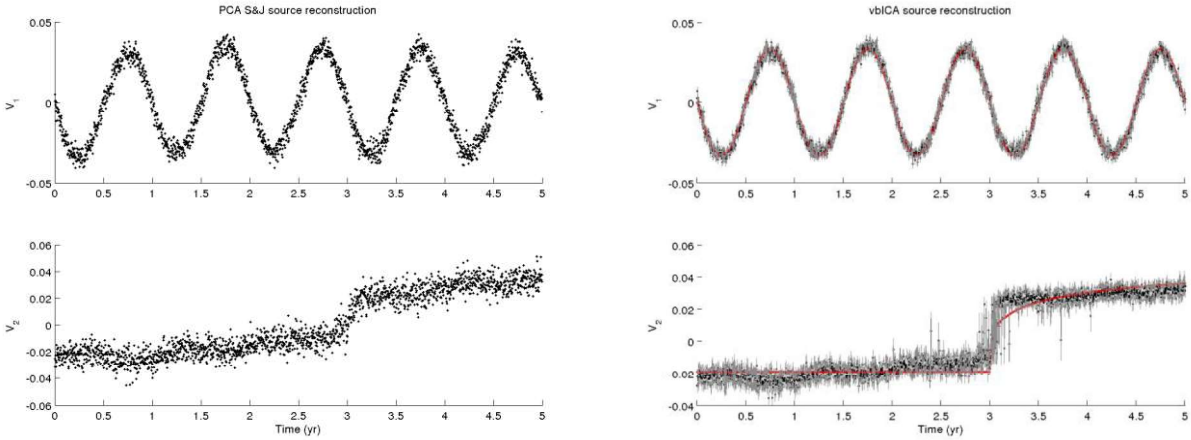
f) vbICA – MD 25%

Figure S5



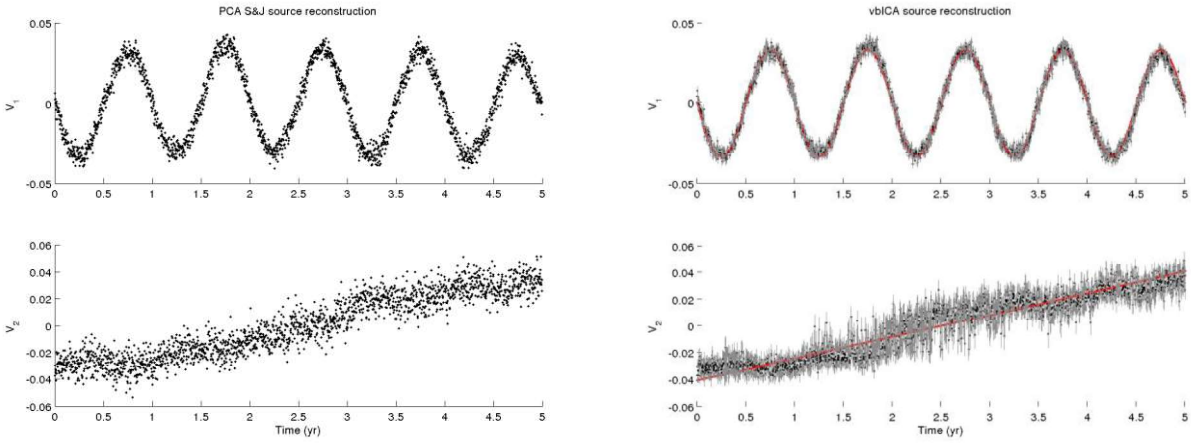
a) PCA – MD 0%

b) vbICA – MD 0%



c) PCA – MD 5%

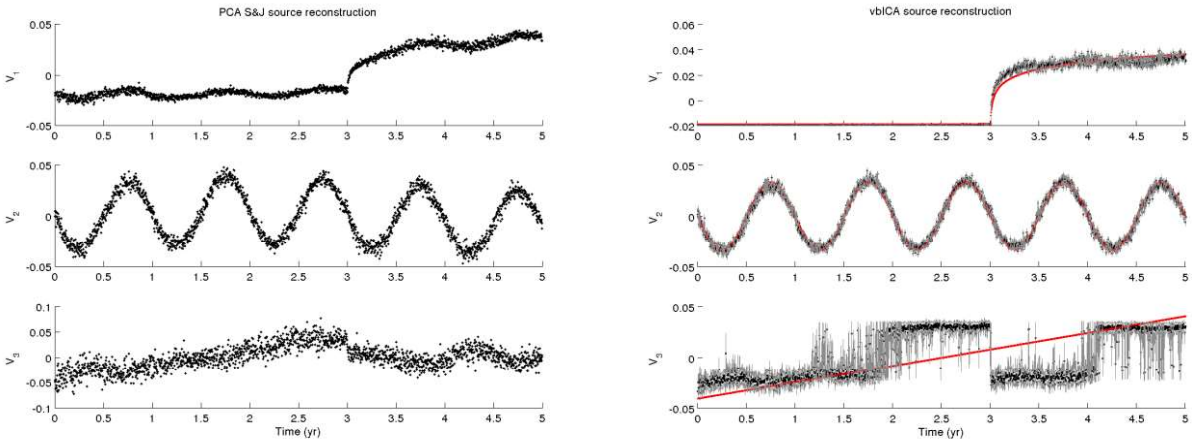
d) vbICA – MD 5%



e) PCA – MD 25%

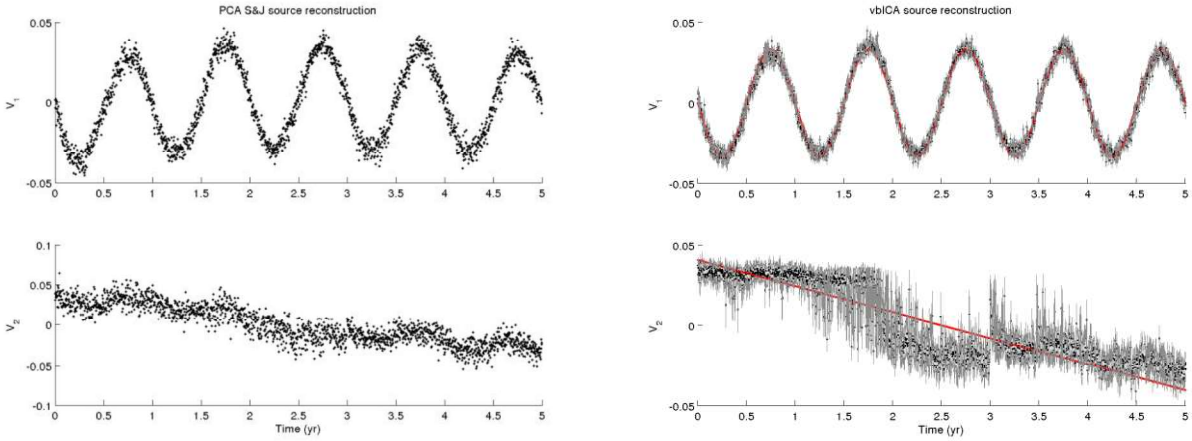
f) vbICA – MD 25%

Figure S6



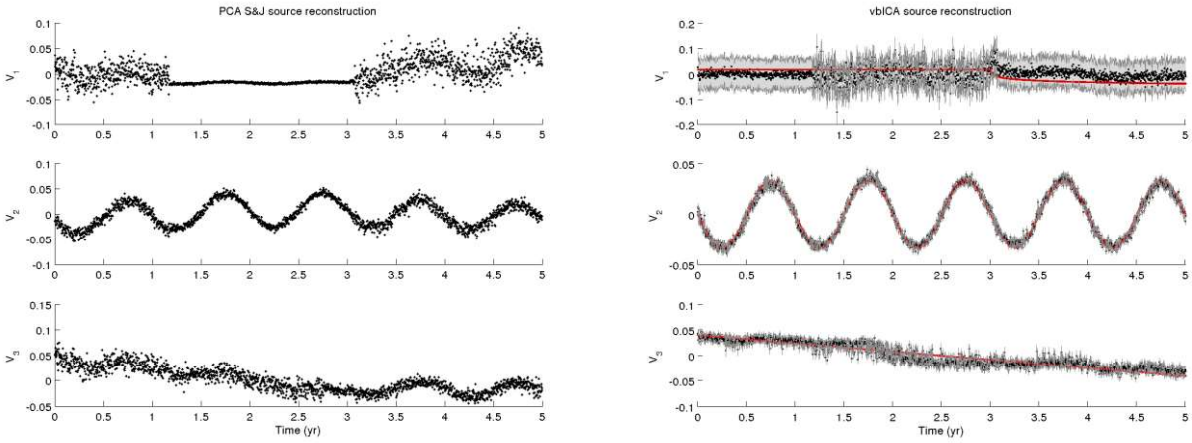
a) PCA – MD 0%

b) vbICA – MD 0%



c) PCA – MD 5%

d) vbICA – MD 5%



e) PCA – MD 25%

f) vbICA – MD 25%

Figure S7

References

Chan K., Lee T.-W. and Sejnowski T.J. (2003) Variational Bayesian Learning of ICA with Missing Data. Neural Comput., 15(8):1991–2011.

Choudrey R.A. (2002) Variational Methods for Bayesian Independent Component Analysis. Pattern analysis and machine learning - robotics research group, University of Oxford. [Available at http://www.robots.ox.ac.uk/~parg/projects/ica/riz/thesis.html, last access 10 June 2015]

Hanks T.C. and H. Kanamori (1979) A Moment Magnitude Scale. J. Geophys. Res., 84(B5), pp. 2348-2350, doi: 10.1029/JB084iB05p02348.

Kanamori H. and E. Brodsky (2004) The physics of earthquakes. Rep. Prog. Phys., 67(1429), doi:10.1088/0034-4885/67/8/R03.

MacKay D.J.C. (1995) Developments in probabilistic modelling with neural networks - ensemble learning. Proceedings of the third Annual Symposium on Neural Networks, Nijmegen, The Netherlands, pp. 191–198, Springer.

Ormerod J.T. and Wand M.P. (2010) Explaining Variational Approximations. Am. Stat., Vol. 64, No. 2, DOI: 10.1198/tast.2010.09058.

Tarantola A. (2005) Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM.