1 Principal component analysis

1.1 Setting

For simplicity, we consider a non-inductive case. Then, for an electric dipole at position $\mathbf{r}_s$ and with dipole moment $\mathbf{p}_s$, the electric field at coordinate $\mathbf{r}$ is obtained from

$$\mathbf{E}^s(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\left(3\frac{\mathbf{p}_s(\mathbf{r}-\mathbf{r}_s)}{r^5}\cdot(\mathbf{r}-\mathbf{r}_s) - \frac{1}{r^3}\mathbf{p}_s\right) \ ,$$

where $r = |\mathbf{r}-\mathbf{r}_s|$. We denote the response due to a unit dipole $\hat{\mathbf{p}}$ as $\hat{\mathbf{E}}(\mathbf{r})$, where $\mathbf{E}^s(\mathbf{r}) = p_s\hat{\mathbf{E}}^s(\mathbf{r})$ and $\mathbf{p} = p\hat{\mathbf{p}}$. Let us distribute a number of S_0 electric dipoles and record the horizontal electric field components with an array of K stations. The total number of recording channels is $N = 2K$.

For the synthetic experiment, we transmit signal with random linear combinations of all sources. Then, at a particular station at position $\mathbf{r}$ the electric field is the superposition of contributions from all source dipoles for a particular random realization i. Hence,

$$\mathbf{E}_i(\mathbf{r}) = \sum_s r_{i,s}\mathbf{E}^s(\mathbf{r}) = \sum_s r_{i,s}p_s\hat{\mathbf{E}}^s(\mathbf{r})$$

where the random number $r_{i,s}$ determines the strength of source s. We collect the responses $E_{n,i}$ at all channels $n = 1, ..., N$ for all realizations $i = 1, ..., I$ in a $N \times I$ data matrix $\mathbf{X}$ as

$$\begin{aligned}
\mathbf{X} &= \begin{bmatrix} E_{1,1} & E_{1,2} & \dots & E_{1,I} \\ E_{2,1} & \ddots & & \\ \vdots & & E_{n,i} & \\ E_{N,1} & & & E_{N,I} \end{bmatrix} \\
&= \begin{bmatrix} \hat{E}_{1,1} & \hat{E}_{1,2} & \dots & \hat{E}_{1,S} \\ \hat{E}_{2,1} & \ddots & & \\ \vdots & & \hat{E}_{n,s} & \\ \hat{E}_{N,1} & & & \hat{E}_{N,S} \end{bmatrix} \begin{bmatrix} p_1 & & & \mathbf{0} \\ & \ddots & & \\ & & p_s & \\ \mathbf{0} & & & p_S \end{bmatrix} \begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,I} \\ r_{2,1} & \ddots & & \\ \vdots & & r_{s,i} & \\ r_{S,1} & & & r_{S,I} \end{bmatrix} \\
&= \hat{\mathbf{E}}\mathbf{p}\mathbf{R}
\end{aligned}$$

1.2 Principal component analysis

The principal components $\mathbf{u}_p$ are defined as linear combinations of the data

$$\mathbf{u}_p = \mathbf{X}\mathbf{a}_p^T$$

under the constraint that $\|\mathbf{a}_p\| = 1$ and $\mathbf{u}_p^T\mathbf{u}_q = 0$ for $p \neq q$. Here, $\mathbf{u}_p$ and and $\mathbf{a}_p$ are $N \times 1$ and $1 \times I$ vectors, respectively. In turn, the data can written as linear combinations of the principal components. In matrix notation, the PCA of the data matrix reads

$$\mathbf{X} = \mathbf{U}\mathbf{A} \ .$$

The dimensions of $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_P]$ and $\mathbf{A} = [\mathbf{a}_1| ...| \mathbf{a}_P]$ are $N \times (P \leq N)$ and $P \times I$, respectively. There is a maximum of $P = N$ principal components, but for low-dimensional data, only the first $P \leq N$ components are required (this means that there are less independent linear combinations than channels). The decomposition of $\mathbf{X}$ is not unique, since $\mathbf{U}\mathbf{A}$ can be replaced with $\tilde{\mathbf{U}}\tilde{\mathbf{A}}$, where $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{B}^{-1}$ and $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{B}$ and $\mathbf{B}$ is any non-singular $P \times P$ matrix. One possible PCA representation is obtained from the singular value decomposition of $\mathbf{X}$, when

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

is truncated at the $P-$th singular value and $\mathbf{A} = \mathbf{S}\mathbf{V}^T$.

A more intuitive explanation of the PCA is as follows: Imagine that the observed fields result from linear combinations of P sources. Then, the maximum number of independent linear combinations is P, and the number of non-zero singular is also P. Therefore, all observed fields reside in a space that is spanned by P orthogonal coordinate axes. These are called the principal components. The current linear combination of sources is reflected in $\mathbf{A}$. Therefore, it is useful to denote the elements of $\mathbf{A}$ as polarization parameters. So, the PCA retrieves the number of independent realizations, and the coordinates of these observations in this coordinate system are $\mathbf{A}$. Other coordinate systems of the same dimension exist also; for instance any rotation is allowed. Therefore the PCA is not unique.

PCA algorithms typically involve re-scaling and leveling of the observations. Here, we use the robust PCA code `robpca`, which is part of the Matlab-based `LIBRA` library. `robpca` finds a robust $N \times 1$ dimensional center $\boldsymbol{\mu}$ of the data and a $N \times P$ dimensional matrix $\mathbf{U}$. The orthogonal columns of $\mathbf{U}$ define a new coordinate system. The polarization parameters $\mathbf{A}$ are the coordinates of the centered data in the new coordinate system, i.e.

$$\mathbf{X} - \mathbf{M} = \mathbf{U}\mathbf{A} \ ,$$

where $\mathbf{M}$ is a matrix containing I replicates of $\boldsymbol{\mu}$ in its columns.

Let us consider the data as linear combinations of a number of contributions from unknown sources

$$\mathbf{X} = \hat{\mathbf{E}}\mathbf{p}\mathbf{R} + \boldsymbol{\delta} \ ,$$

where δ represents some noise. There is only a limited number of independent source configurations that can be discriminated by the data. To investigate this, we apply a PCA analysis on $\mathbf{X}$ and keep the P dominant components, i.e.

$$\mathbf{X} - \mathbf{M} = \mathbf{U}\mathbf{A} = \hat{\mathbf{E}}\mathbf{p}\mathbf{R} + \boldsymbol{\delta} \ .$$

Recall that $\mathbf{A}$ contains the polarization parameters for each single event. One particular set of P independent realizations would correspond to $\mathbf{A} = \mathbf{I}_{P\times P}$ being a unity matrix. Ignoring $\mathbf{M}$ and $\boldsymbol{\delta}$ for the moment, we aim at determining the source parameters $\mathbf{S} = \mathbf{p}\mathbf{R}$ for this set of realizations for an assummed distribution of sources that generate the fields $\hat{\mathbf{E}}$ at the actual receiver locations. Let this problem be mixed-determined in the sense that we allow for more unknown sources S than independent components P. Then, the least squares estimate for the source parameters is

$$\mathbf{S}^{est} = (\hat{\mathbf{E}}^T\hat{\mathbf{E}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{E}^T}\mathbf{U}$$

where λ is a regularization parameter. Each of the columns of $\mathbf{S}^{est}$ then corresponds to the moments of all sources S which, when superposed, give rise to the fields given by the columns of $\mathbf{U}$.

Example: Field of an electric dipol in an insulating full-space

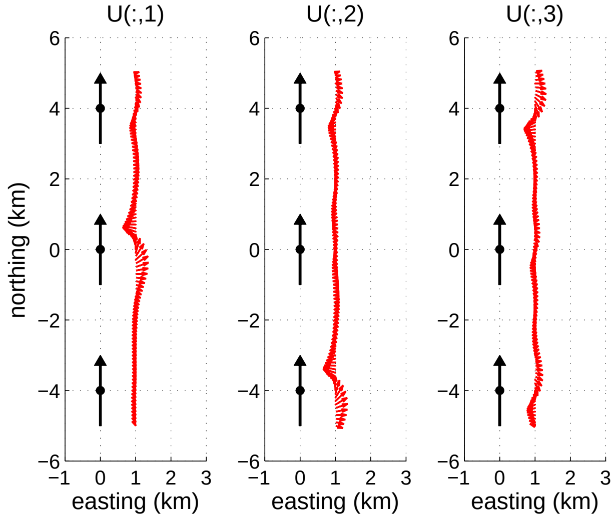

Figure 1: Black arrows indicate position and orientation of three true source dipoles. Thousand random linear combinations of the fields generated by these sources have been collected. The PCA yields three independent source configurations, which correspond to three particular but unknown source combinations. The electric is plotted for each receiver and each of the principal components (corresponding to the columns of $\mathbf{U}$).

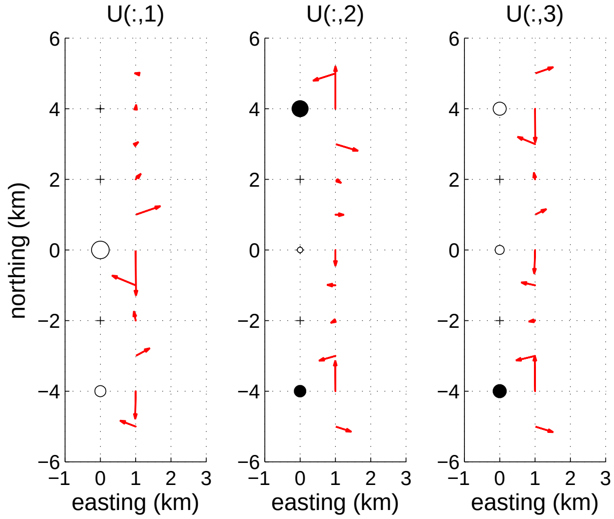

Figure 2: Reconstruction of source moments required to explain the principal components. Here, only a subset of 11 stations has been used. For the reconstruction, we place the three sources exactly at the position of the true sources. Dipole moments correspond to the size of circles; white and black correspond to negative and positive moments. Without noise, the reconstruction is perfect.

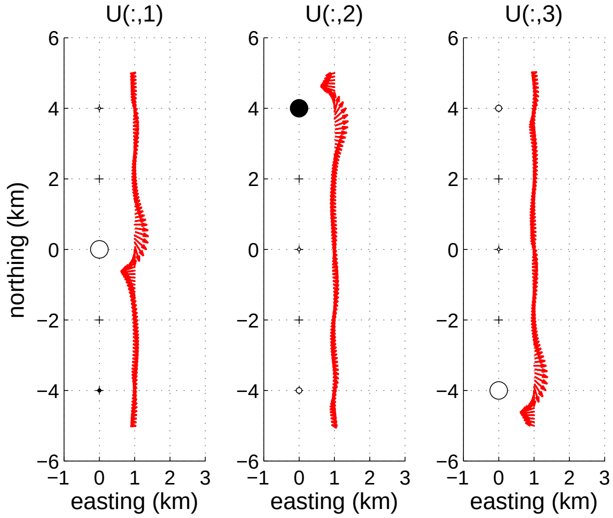

Figure 3: Same as before, but 100 receivers have been used in the analysis, and 5 unknown sources were considered. The reconstruction of source moments yields a linear combination of three sources (those which are at the position of the true sources), whereas the estimated dipole moments for remaining two is zero.

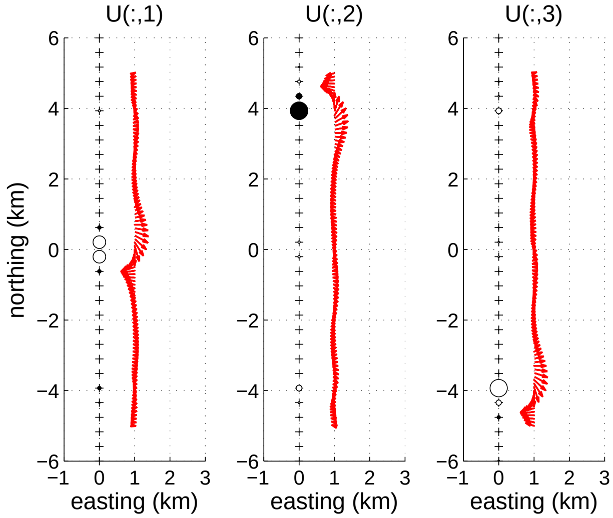

Figure 4: Same as before, but 30 source dipoles are included to estimate the solution. None of the saught dipoles is at the location of a true source. The solution is regularized.

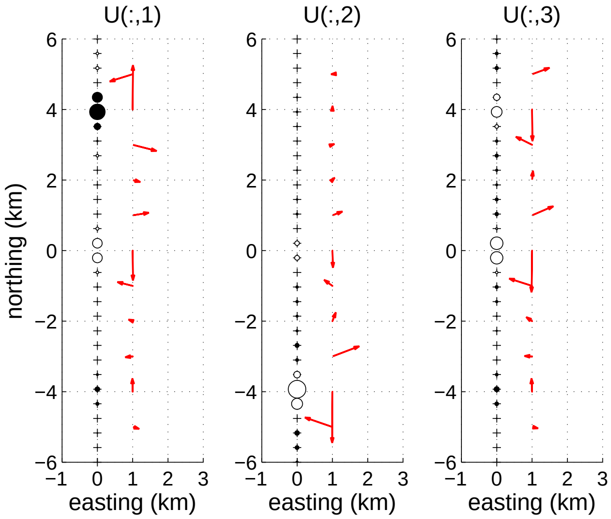

Figure 5: Same as before, but only 11 receivers were used.

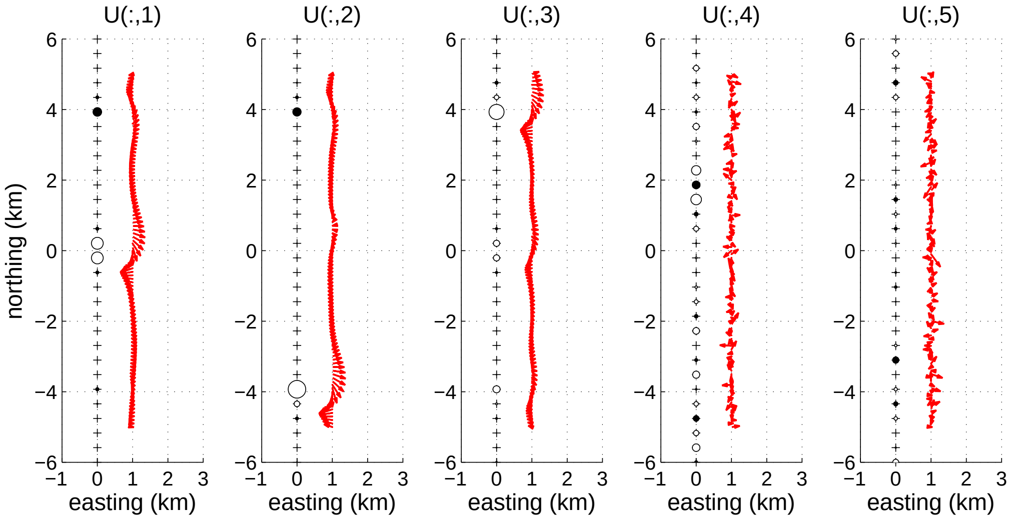

Figure 6: Same as before, but data were afflicted with random noise. The 4th and 5th principal components are unimportant but are displayed for completeness. They contain the noise.