doi:10.1017/S143192761005734X
Microsc. Microanal. 16 (Suppl 2), 2010

Probabilistic Principal Component Analysis using Expectation Maximization (PPCA-EM) for Analyzing 3D Volumes with Missing Data

Lingbo Yu,*,** Robert R. Snapp,** Teresa Ruiz,* and Michael Radermacher*,**

* Department of Molecular Physiology and Biophysics, University of Vermont, Burlington, Vermont, VT 05405
** Department of Computer Science, University of Vermont, Burlington, Vermont, VT 05405

There are different techniques to obtain 3D structures of macromolecular assemblies from electron micrographs: electron tomography with either single-axis, dual-axes or conical tilting [1,2], random conical reconstruction [3,4], angular reconstitution [5], and orthogonal tilting reconstruction [6]. The techniques including tilting are the most suitable for investigating structures of heterogeneous samples. However, most tilting techniques leave out areas of missing data in the 3D reconstructions because the tilt angle range is limited in the electron microscope. If the volumes are combined correctly, a volume with no missing data or much less missing data can be achieved. This requires correct identification of the volumes representing the same structure. However, missing data present an obstacle to analyzing variations between 3D volumes since they cause artifacts in the 3D reconstructions. Especially when the data is missing in different orientations, artifacts can be easily misinterpreted as structural differences.

Here we present an algorithm, Probabilistic Principal Component Analysis using Expectation maximization algorithm (PPCA-EM), which was adopted from Tipping and Bishop's framework [7,8], and which can perform principal component analysis (PCA) on a set of 3D volumes with arbitrary missing data. Like traditional PCA, PPCA-EM not only reduces the dimensionality of the data which reduces the complexity of any subsequent classification, but also reduces the noise, thus increasing the robustness of the classification. Unlike the traditional PCA, PPCA-EM estimates the missing observations as well. This becomes highly valuable when no clearly defined classes exist but the data shows continuous variations that prevent commonly used averaging techniques from being applied.

The relation between the principal components and the original volumes is modeled as:

$$\begin{pmatrix} \mathbf{t}_i^{(p)} \\ \mathbf{t}_i^{(m)} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_i^{(p)} \\ \mathbf{W}_i^{(m)} \end{pmatrix} \mathbf{x}_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i, i = 1,2,\ldots n, \quad (1)$$

where $\mathbf{t}_i^{(p)}$ and $\mathbf{t}_i^{(m)}$ are the present part and the missing part of the ith volume, $\mathbf{x}_i$ represents the ith volume in a lower dimensional subspace, $\boldsymbol{\mu}$ is mean of the data set containing n 3D volumes, and $\boldsymbol{\varepsilon}_i$ is the residual. $\mathbf{W}$ is a transform matrix, whose rows separate into $\mathbf{W}_i^{(p)}$ and $\mathbf{W}_i^{(m)}$ respectively, and whose columns $\mathbf{w}_j$ correspond to features or eigenvectors. The underlying probabilistic models for traditional PCA are applied, which are $\mathbf{x}_i \sim N(0,\mathbf{I})$ and $\boldsymbol{\varepsilon}_i \sim N(0,\sigma^2\mathbf{I})$. Incorporated in the algorithm is an additional index vector $\boldsymbol{\rho}_i$ used to separate the missing part from the present part. $\boldsymbol{\rho}_i$ is either provided as part of the reconstruction algorithm, or can be determined from the reconstruction geometry.

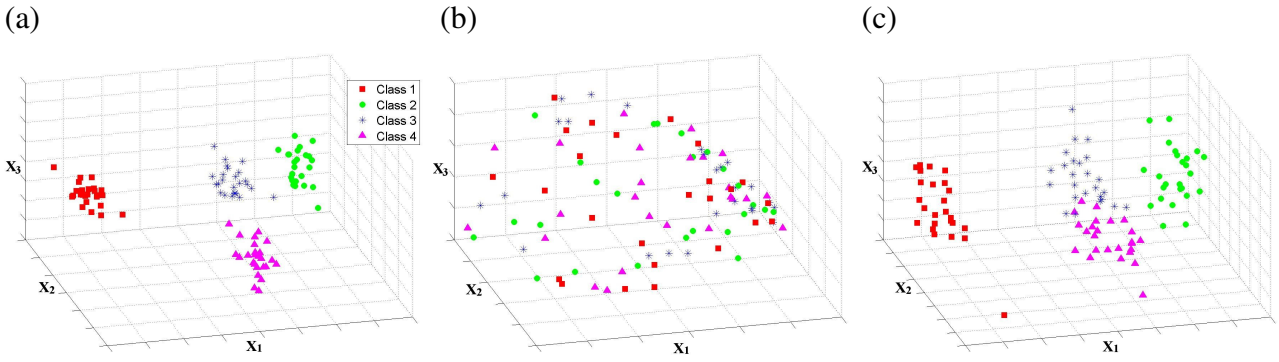
An expectation maximization algorithm is applied to solve the model, Eq. 1. In the E-step, hidden variables, $\mathbf{x}_i$ and $\mathbf{t}_i^{(m)}$, are estimated based on the present part $\mathbf{t}_i^{(p)}$, which eliminates the artifacts caused by the missing data. In the M-step, the transform matrix $\mathbf{W}$ is updated. The E-step and M-step are iterated until a predefined convergence criterion is met or a maximum number of iterations is reached.

The algorithm has been tested extensively on simulated 3D data. 2700 tests were carried out on volume sets containing four slightly different conformations of a macromolecule, with variations of a size similar to variations observed in experimental data. Different sets were created by varying the signal-to-noise ratio (SNR), the percentage of missing data and the number of volumes in a set. The test results clearly show the strength and limitation of the algorithm. One of the test sets is shown in (Fig. 1), containing 100 3D polar Fourier volumes at SNR 0.5 with 30% data missing. The separation of the points in Fig.1c shows that the algorithm can correctly group the volumes regardless of the missing data. First tests on real experimental images have been successful and show that the assumption of a Gaussian noise model is a reasonable approximation for real data.

References

[1] W. Hoppe, et al., Z. Naturforsch Section a, vol. 31a (1976) 645.
[2] M. Radermacher and W. Hoppe, Proc. 9th Int. Congr. Electron Microsc., vol. 1 (1978) 218.
[3] M. Radermacher, et al., Journal of Microscopy, vol. 141 (1986) Rp1.
[4] M. Radermacher, et al., Journal of Microscopy, vol. 146 (1987) 113.
[5] Van Heel, Ultramicroscopy, vol. 21 (1987) 111.
[6] A.E. Leschziner and E. Nogales, J. Struct Biol, vol. 153 (2006) 284.
[7] M.E. Tipping and C.M. Bishop, J. of the Royal Stat. Soc.: Series B: Stat. Methodology, vol. 61, (1999) 611.
[8] S. Roweis, Neural Information Processing Systems (1997) 626.
[9] This work was supported by NIH grant RO1 GM078202 (to M.R.).

Figure 1.



Scatter plots of the real part of the first three principal components. Symbols correspond to the true classes of each volume. (a) Data set with no missing data, standard PCA. The four classes are clearly separated. (b) 30% missing data, standard PCA. The results are dominated by the missing data. (c) 30% missing data, determined with PPCA-EM, showing the correct clustering of the volumes.

https://doi.org/10.1017/S143192761005734X Published online by Cambridge University Press