Accounting for Probe-level Noise in Principal Component Analysis of Microarray Data

Guido Sanguinetti, Marta Milo, Magnus Rattray and Neil D. Lawrence

http://www.bioinf.man.ac.uk/resources/puma/

The University Of Sheffield.

Machine Learning Group
Computer Science
University of Sheffield, U.K.

The University of Manchester

Artificial Intelligence Group
Computer Science
University of Manchester, U.K.

Overview

- Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques for the analysis of high dimensional datasets.
- In its standard form, PCA assumes all data points to be i.i.d. and corrupted by i.i.d. noise.
- This is often an unreasonable assumption when dealing with microarray data, where different genes and different conditions have different levels of experimental and biological noise.
- We propose a new model-based approach to PCA that takes into account the variances associated with each gene in each experiment [8].
- The model provides signifi cantly better results than standard PCA, and avoids arbitrary manipulations such as setting cut-offs on expression levels.
- We demonstrate how the model can be used to denoise a data set, leading to improved expression profi les and tighter clusterings.

Motivation

- PCA is one of the most popular techniques for extracting information from high dimensional datasets.
- It is very popular in microarray data analysis, where the principal components are interpreted as the (few) physiological processes driving the variability in the data set.
- PCA makes two implicit assumptions: the fi rst is that the data is normally distributed, the second is that the uncertainty associated with each gene under each condition is constant. This second assumption is often very unrealistic in biological data.
- Traditionally, this problem has been avoided by introducing cut-offs at the preprocessing stage. This involves a large degree of arbitrariness in selecting the cut-off and potentially throws away useful information about low expressed genes.
- Recent techniques allow to estimate credibility intervals for each gene expression in each time point in a microarray experiment ([4],[5],[3] and [6]).
- We seek to avoid ad hoc manipulations and propagate the uncertainties through a probabilistic model as a principled way of avoiding the problems inherent with PCA.

Probabilistic PCA

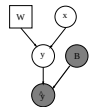
- Our model is a generalisation of the Probabilistic PCA algorithm ([9]).
- This is a latent variable model where each d-dimensional data point $\mathbf{y}_n$ can be reconstructed from a q-dimensional latent point $\mathbf{x}_n$ via a linear transformation W and a corrupting noise vector ϵ_n.
$$\mathbf{y}_n = W\mathbf{x}_n + \mu + \epsilon_n.$$
- The noise vector is assumed to come from a spherical Gaussian distribution $\epsilon_n \sim N\left(0, \sigma^2 I\right)$, which implies a likelihood
$$\mathbf{y}_n|\mathbf{x}_n \sim N\left(W\mathbf{x}_n + \mu, \sigma^2 I\right).$$
- Integrating over the latent variable $\mathbf{x}$, one obtains the marginal likelihood
$$\mathbf{y}_n \sim N\left(\mu, WW^T + \sigma^2 I\right). \quad (1)$$
- Maximising the likelihood one obtains an estimate of the matrix W s.t. its columns span the principal subspace of the data space.

Limits of Probabilistic PCA and Factor Analysis

- Probabilistic PCA assumes the data to be i.i.d. and explains all noise effects with an indiscriminate spherical Gaussian.
- A more flexible model is Factor Analysis [1], which allows each dimension to have a different noise variance
$$\epsilon \sim N\left(0, B^{-1}\right),$$
where B is a diagonal matrix containing the individual precisions.
- Factor Analysis however is still not flexible enough to handle common microarray situations, where errors can vary greatly between genes and between different conditions for the same gene. We therefore need a model which can handle non-i.i.d. data.

Propagating Uncertainty

- We consider a modifi ed model where the "true" expression levels are i.i.d. data but each measurement is corrupted by white noise with a different (known) variance. We assume that these variances have been measured at a preprocessing stage, using one of the various existing methods for obtaining credibility intervals from microarray experiments.
- Take $\mathbf{y}_n$ to be a d-dimensional vector which represents the true log expression level associated with the n-th gene under d different conditions. Rather than observing $\mathbf{y}_n$ directly we assume that we observe a corrupted form $\hat{\mathbf{y}}_n$
$$\hat{\mathbf{y}}_n = \mathbf{y}_n + \boldsymbol{\nu}_n \quad (2)$$
where $\boldsymbol{\nu}_n$ is noise which is distributed as
$$\boldsymbol{\nu}_n \sim N\left(0, B_n^{-1}\right).$$
Here, B_n is a diagonal matrix whose ith diagonal element is given by β_{ni} which is the precision associated with the i-th experiment for the n-th gene. This precision can be obtained through one of the probabilistic analysis methods mentioned above. A graphical model represention of our model is given below.



Graphical representation of the noisy PPCA model.

- We will assume a probabilistic PCA model as the marginal distribution for the true expression level $\mathbf{y}_n$, as given in (1), and obtain,
$$\hat{\mathbf{y}}_n|\mathbf{x}_n \sim N\left(W\mathbf{x}_n + \boldsymbol{\mu}, \sigma^2 I + B_n^{-1}\right). \quad (3)$$
We denote collectively $A_n = \sigma^2 I + B_n^{-1}$ and using $\mathbf{x}_n \sim N(0, I)$, we have the following marginalised likelihood,
$$\hat{\mathbf{y}}_n \sim N\left(\boldsymbol{\mu}, WW^{\mathrm{T}} + A_n\right). \quad (4)$$
- The corrupted data is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $C_n \doteq WW^{\mathrm{T}} + A_n$. Notice however that, as the data is not i.i.d., the maximum likelihood estimator of the mean vector $\boldsymbol{\mu}$ does no longer coincide with the empirical mean, but must be learnt alongside the other parameters.

Efficient Likelihood Optimisation

- Given the marginal likelihood of equation (4), we can optimise the parameters through a non-linear optimisation such as scaled conjugate gradients.
- This is generally computationally unfeasible for large data sets. A more effi cient algorithm can be obtained through an expectation maximisation (EM) approach [2].
- Generally EM algorithms lead to a simplifi ed optimisation problem (the M-step) by incorporating an additional step (the E-step).
- For our corrupted data PCA model this additional step is the computation of the posterior distribution for the latent space. This posterior is obtained through Bayes' theorem
$$\mathbf{x}_n|\hat{\mathbf{y}}_n \sim N\left(M_n W^{\mathrm{T}} A_n^{-1}\left(\hat{\mathbf{y}}_n - \boldsymbol{\mu}\right), M_n\right) \quad (5)$$
where we have defi ned
$$M_n = \left[W^{\mathrm{T}} A_n^{-1} W + I\right]^{-1}.$$
- The EM algorithm then iteratively updates the posterior distribution (E-step) and maximises a lower bound on the likelihood with respect to the model parameters (M-step). The algorithm provably converges to a maximum of the likelihood.

E-step

- The lower bound on the likelihood is given by
$$\mathcal{L}_c = -\frac{1}{2}\sum_{n=1}^{N}\log|A_n| + \sum_{n=1}^{N}\mathrm{tr}\left(\left\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right\rangle\right) + \sum_{n=1}^{N}\left(\hat{\mathbf{y}}_n - \boldsymbol{\mu}\right)^{\mathrm{T}} A_n^{-1}\left(\hat{\mathbf{y}}_n - \boldsymbol{\mu}\right)$$
$$-2\sum_{n=1}^{N}\left\langle \mathbf{x}_n\right\rangle^{\mathrm{T}} W^{\mathrm{T}} A_n^{-1}\left(\hat{\mathbf{y}}_n - \boldsymbol{\mu}\right) + \sum_{n=1}^{N}\mathrm{tr}\left(W^{\mathrm{T}} A_n^{-1} W\left\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right\rangle\right), \quad (6)$$
where the notation $\langle\cdot\rangle$ denotes the expectation under the posterior distribution over $\mathbf{x}_n$ (5).
- The E-step evaluates these expectations from the suffi cient statistics of the posterior distribution over $\mathbf{x}_n$.

M-step

- Taking the gradients of eq. (6), one obtains fi xed point equations for W and $\boldsymbol{\mu}$ which give an approximate M-step
$$\boldsymbol{\mu} = \left(\sum_{n=1}^{N} A_n^{-1}\right)^{-1}\sum_{n=1}^{N} A_n^{-1}\left(\hat{\mathbf{y}}_n - W\left\langle \mathbf{x}_n\right\rangle\right),$$
$$W_{ji} = \sum_{n=1}^{N} H_{jn}\left(L_i\right)_{ni}^{-1}, \quad (7)$$
where A_{nj} is the j-th diagonal element of A_n and we have introduced the two matrices
$$H = \sum_{n=1}^{N} A_n^{-1}\left(\hat{\mathbf{y}}_n - \tilde{\boldsymbol{\mu}}\right)\left\langle \mathbf{x}_n\right\rangle^{\mathrm{T}},$$
$$L_j = \sum_{n=1}^{N} A_{nj}^{-1}\left\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right\rangle.$$
- These update equations can be iterated to fi nd the maximum likelihood solution with respect to W and $\boldsymbol{\mu}$.
- A fi xed point equation for σ^2 (which accounts for any residual variance) cannot be obtained as the gradient with respect to σ^2 is not linear. An effi cient update for σ^2 can be obtained by using Newton's method.

Number of Principal Components

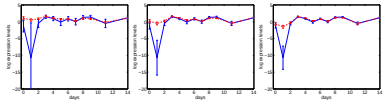
- The usual approach when implementing PCA for microarray data is to retain a reduced number of principal directions, q, and project the log expression levels along these directions before further processing.
- In general, the number of principal components retained is pre-determined according to the specifi c problem under consideration.
- In our model, however, the estimate of the noise allows to evaluate the statistical signifi cance of a direction.
- Therefore, we automatically obtain the maximum number of principal components that can be retained.

Data sets

- Data set consisted of a temporal assay of Affymetrix GeneChip arrays that measured the gene expression profi les of a conditionally immortal cell line, UB/OC-1, from mouse cochlear epithelial cells at embryonic day 13.5 (E13.5), across 14 days of differentiation [7].
- Of particular interest in this study is the identifi cation of targets regulated by the transcription factor gata-3, which is essential for normal inner ear development .
- In vivo the expression values of gata-3 are low before day 4 when they start to rise. They peak at day 8-9 and after a couple of days the expression level decreases again to then stabilize around a constant value.
- The raw data was processed using a modifi ed version of the gMOS algorithm [6][5].

Profile Reconstruction

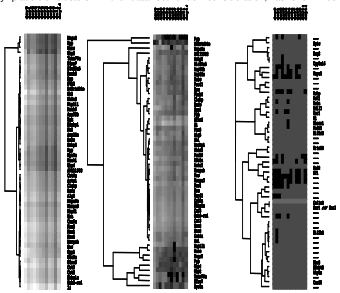
- The estimates of the parameters, together with the expectations of the hidden variables $\mathbf{x}_n$, can be used in equation (3) to obtain estimates of the true gene expression levels and their covariances, given by
$$\bar{\mathbf{y}}_n = W\left\langle \mathbf{x}_n\right\rangle + \boldsymbol{\mu}$$
$$\Sigma_n = W\left[\left\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right\rangle - \left\langle \mathbf{x}_n\right\rangle\left\langle \mathbf{x}_n^{\mathrm{T}}\right\rangle\right]W^{\mathrm{T}} + \sigma^2 I.$$
- We can then obtain an estimated profi le for the "true" expression levels.
- To show the effect of the uncertainty in the data,we modelled the data three times, artifi cially reducing the variances by factors 1, 4 and 9. The corrected expression profi les are shown below.
- Note that as the uncertainty in the original profi le is decreased the corrected profi le tends to stay closer to its original course. As can be seen from the plots, any point with large associated uncertainty (such as the day 1 point for the gata-3 profi le) can be signifi cantly changed and this can lead to a large decrease in the associated uncertainty.



Corrected profi le (thick dashed line) and original profi le (thin solid line) for the gata-3 gene (a transcription factor) *left*: corrected profi le based using the original uncertainties; *middle*: corrected profi le with the uncertainty halved and *right*: corrected profi le with a third of the original uncertainty.

Clustering

- Clustering is a widely used technique for summarising expression levels obtained from microarray data and as an exploratory technique for fi nding functional analogues.
- One suggested use of PCA in microarray analysis is as a preprocessing step before cluster analysis. The use of PCA before clustering can be justifi ed by the fact that the larger principal components are expected to capture the structure in the data set.
- However, standard PCA does not always improve the clustering but often degrades it, since the dominant components, which contain most of the variation in the data, are highly influenced by the very noisy data points.
- By accounting for the variance in the log expression levels, our algorithm automatically downweights noisy values and ensures that the components we extract accurately reflect the structure of the data.
- The clustering is further improved when performed on the denoised reconstructed profi les, as these are the best estimates of the true profi les. This leads to much tighter and biologically plausible clusters in the data set under consideration, as shown below.



Hierarchical clustering of microarray data *left*: the top 50 genes in the second principal component obtained using our model (denoised profi les);*middle*: the top 50 genes in the second principal component obtained using our model (original profi les) and *right*: the top 50 genes in the second principal component obtained by standard PCA. Clustering was performed using the GeneCluster software from the Eisen Lab.

Acknowledgements

GS, MR and NL gratefully acknowledge support from a BBSRC award *"Improved processing of microarray data with probabilistic models"*. MM is supported by an Advanced Training Fellowship from the Wellcome Trust.
The software used can be downloaded from http://www.bioinf.man.ac.uk/resources/puma/

References

[1] David J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffi n & Co. Ltd, London, 1987.

[2] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

[3] Anne-Mette K. Hein, Sylvia Richardson, H. C. Causton, Graeme K. Ambler, and Peter J. Green. BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics*, 2005.

[4] Neil D. Lawrence, Marta Milo, Mahesan Niranjan, Penny Rashbass, and Stephan Soullier. Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics*, 20(4):518–526, 2004.

[5] Xuejun Liu, Marta Milo, Neil D. Lawrence, and Magnus Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.

[6] Marta Milo, Alireza Fazeli, Mahesan Niranjan, and Neil D. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transations*, 31(6):1510–1512, 2003.

[7] Marcelo N. Rivolta, A. Halsall, C. Johnson, M. Tones, and Matthew C. Holley. Genetic profi ling of functionally related groups of genes during conditional differentiation of a mammalian cochlear hair cell line. *Genome Research*, 12(7):1091–1099, 2002.

[8] Guido Sanguinetti, Marta Milo, Magnus Rattray, and Neil D. Lawrence. Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 2005.

[9] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999.