Bayesian Extreme Components Analysis

Yutian Chen **Max Welling**
Bren School of Information and Computer Science
University of California, Irvine
{yutianc,welling}@ics.uci.edu

Abstract

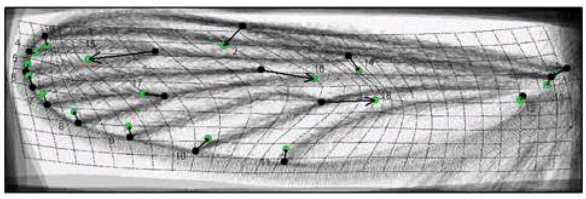
Extreme Components Analysis (XCA) is a statistical method based on a single eigenvalue decomposition to recover the optimal combination of principal and minor components in the data. Unfortunately, minor components are notoriously sensitive to overfitting when the number of data items is small relative to the number of attributes. We present a Bayesian extension of XCA by introducing a conjugate prior for the parameters of the XCA model. This Bayesian-XCA is shown to outperform plain vanilla XCA as well as Bayesian-PCA and XCA based on a frequentist correction to the sample spectrum. Moreover, we show that minor components are only picked when they represent genuine constraints in the data, even for very small sample sizes. An extension to mixtures of Bayesian XCA models is also explored.

1 Introduction

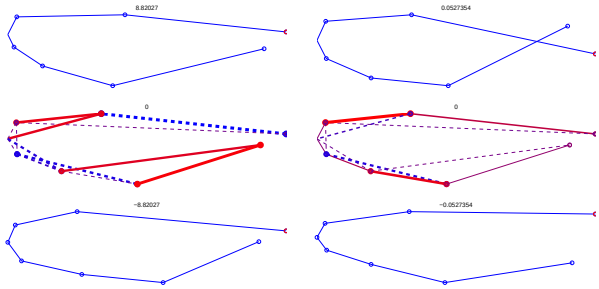
Principal components analysis (PCA) is perhaps the most widely used algorithm in the fields of statistics, machine learning and data mining. It enjoys a number of desirable properties such as optimal reconstruction of the original signal in the L_2 norm and retaining the maximal variance directions in the data. Despite this, there are many examples where it is not the principal components (PCs) that convey the important information, but rather the directions of small variance, or *minor components* (MCs). One can think of minor components as properties that are conserved in the data, i.e. constraints.

As a motivating example we consider the landmark measurements on the edge of a mosquito wing[1], see Figure 1(a). In our experiments we have sub-sampled a number of these landmarks located on the edge of the wing and computed the relative distances between these landmarks and some of their nearest neighbors. We have used distances as features because they remove translational and rotational degrees of freedom (for more details see section 5.1.)

In Figure 1(b) we have plotted the deformation of the geometry of the landmark configuration as we change the coefficient for the principal component or the minor component of the data. One can observe that the first PC corresponds to a shift of the landmarks over the edge of the wing, but keeping the wing shape mostly invariant, indicating that the location where the veins of the wing intersect with the boundary is highly variable across mosquitos. Looking at the minor component of the data we see that it (when varied) would drastically change the shape of the wing, in particular the part where the wing is attached to the mosquito's body. Since minor components express variability which is absent in the data it implies that this type of shape change is highly unlikely in the mosquito population. We argue that these "conservation laws of biological evolution" are of more scientific interest



(a) Mosquito Wing Landmarks (dots). Landmarks are placed where the veins intersect with the wing's outer boundary.



(b) Deformation of the geometry of the landmark configurations along the first PC (left) and first MC (right). The middle two figures show the mean positions of the landmarks. The width of the lines represent the weight associated with the features (i.e. distances) for the respective PC/MC. Plots in the left column show the wing as we deform it by varying the coefficient associated with first PC from positive (top) to negative (bottom) and similarly for the MC on the right.

Figure 1: Mosquito Wing Landmarks

[1] Obtained from *http://life.bio.sunysb.edu/morph/index.html*

than the directions of high variability.

More generally, for an arbitrary dataset we would like to reliably determine which linear subspace constitutes an optimal description of the data. A statistical technique called eXtreme Components Analysis (XCA) was introduced in [Welling *et al.*, 2003] to determine the optimal combination of principal and minor components automatically from data. In a statistical sense, it is very difficult to reliably estimate minor components from data if the number of data instances is relatively small compared to the number of attributes. This effect is illustrated in Figure 2 where we plot the sample spectrum computed from a multivariate normal distribution with unit variance in all directions. Even though this data should clearly have no preference for either principal or minor components, sample fluctuations always create artificial low variance directions which correspond to under-sampled directions in space. Since an eigenvalue decomposition searches for these directions it is highly prone to over-fitting to this type of sampling noise.

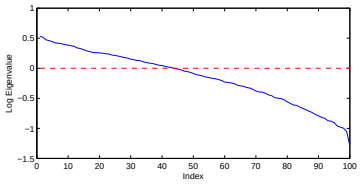


Figure 2: The sample log-spectrum (solid) computed from 200 samples drawn from a 100-D normal distribution with unit variance in all directions. The true log-spectrum is a straight line (dashed).

The contribution of this paper is to largely resolve this issue by introducing a conjugate prior to the parameters of the XCA model which in turn regularizes the eigenvalue decomposition. The effect of this is that minor components will only be incorporated in the model if they represent genuine constraints in the data and not merely under-sampled directions of space.

In our empirical evaluation we show that Bayesian XCA performs at least as good, and often better than 1) XCA based on frequentist shrinkage estimates of the sample covariance, 2) Bayesian PCA and 3) plain vanilla XCA. These results still hold true for an extension to mixtures of Bayesian XCA models that we discuss as well.

2 Extreme Components Analysis

XCA models a data cloud probabilistically by starting with an isotropic Gaussian pdf and either stretching in certain directions (the PCs) or contracting it in certain direction (the MCs). Denote these directions with w_i (organized as rows of W).

$$z_1 = Wx \quad z_1 \sim \mathcal{N}[0, I_d], \quad z_2 = Vx \quad z_2 \sim \mathcal{N}[0, vI_{(D-d)}] \tag{1}$$

where z_1 and z_2 represent independent subspaces and the rows of V form an orthonormal basis in the orthogonal complement of the space spanned by W. The distribution of XCA is normal with inverse covariance matrix given by

$$C_{XCA}^{-1} = \frac{V^T V}{v} + W^T W. \tag{2}$$

We can decompose W^T using an SVD as follows,

$$W^T = UL^{-1/2}R \tag{3}$$

where $U \in \mathcal{R}_{D\times d}$,$R \in \mathcal{R}_{d\times d}$ satisfying $U^T U = I_d$ and $R^T R = I_d$. L is diagonal matrix with elements $\{l_i\}$. One can show that the variance of data in the direction u_i (ith column of U) is given by l_i and in the remaining directions by v. Data is elongated in directions where $l_i > v$ and contracted in directions where $l_i < v$. When all the l_i's are larger than v, XCA is equivalent to probabilistic PCA (PPCA) [Roweis, 1998; Tipping and Bishop, 1999b]. On the other hand, when all the l_i's are smaller than v, XCA is equivalent to probabilistic MCA (PMCA)[Williams and Agakov, 2002].

3 Bayesian XCA

Since XCA is more flexible than PCA and MCA, it will always have equal or higher likelihood on the training set. However, XCA can be sensitive to overfitting. The selection of PCs and MCs depends on the estimation of the sample spectrum, but both the variance as well as the bias of ordered eigenvalues of the sample covariance matrix increase as the ratio of the number of data instances to the number of attributes decreases. In the extreme case where the number of data instances is smaller than the number of attributes, the smallest eigenvalues are 0 and XCA will always pick these minor components. This results in a positive infinite log-likelihood of training data but a negative infinite log-likelihood of test data. Thus, regularization is necessary for XCA when the size of the training dataset is small compared to the number of attributes.

3.1 A Prior for XCA:

In this paper, we use a prior for the parameters of XCA to regularize the estimation of the eigenvalues. When the number of data cases increases, the effect of the prior will automatically diminish. We are aware of two approaches to Bayesian formulations of PPCA [Bishop, 1999; Minka, 2000]. Unlike [Bishop, 1999], the method of [Minka, 2000] can be extended to XCA as we will describe below in more detail[2].

The probability of the dataset $\mathcal{D}$ given the covariance matrix C and mean m is under the XCA model is,

$$P_{XCA}(\mathcal{D}|C,m) = (2\pi)^{-ND/2}|C^{-1}|^{N/2}\exp(-\frac{1}{2}tr(C^{-1}S)) \tag{4}$$

where $S = \sum_n (x_n - m)(x_n - m)^T$, $C^{-1} = \frac{1}{v}(I - UU^T) + UL^{-1}U^T$, N is the number of data points and D the dimensionality of data.

We use a conjugate prior for (U, L, R, v, m) controlled by two hyper-parameters α, β, similar to those used in [Minka,

[2]We thank T. Minka for his prompt responses to our questions about his method. The prior in this paper is similar to that in an updated version of [Minka, 2000].

2000].

$$p(U,L,R,v,m) \propto |C^{-1}|^{\alpha/2} \exp(-\frac{\alpha\beta}{2} tr(C^{-1})) \quad (5)$$

$$\propto |L|^{-\alpha/2} v^{-(D-d)\alpha/2} \exp(-\frac{\alpha\beta}{2} tr(L^{-1})) \exp(-\frac{\alpha\beta(D-d)}{2v})$$

The prior for m is constant in an area large enough for the problem.This distribution factors into separate terms for (U, L, R, v):

$$p(U,L,R,v,m) = p(v)p(U)p(R)\prod_{i=1}^{d} p(l_i) \quad (6)$$

$$p(v) \sim \chi^{-2}(\alpha(D-d)-2, \alpha\beta(D-d)) =$$

$$\frac{\exp(-\frac{\alpha\beta(D-d)}{2v})}{\Gamma(\alpha(D-d)/2-1)v} \left(\frac{\alpha\beta(D-d)}{2v}\right)^{\alpha(D-d)/2-1}$$

$$p(U)p(R)p(m) = constant$$

$$p(l_i) \sim$$

$$\chi^{-2}(\alpha-2, \alpha\beta) = \frac{1}{\Gamma(\alpha/2-1)l_i} \left(\frac{\alpha\beta}{2l_i}\right)^{\alpha/2-1} \exp(-\frac{\alpha\beta}{2l_i})$$

The priors for l_i and v are χ^{-2} distributed with mode β while for other irrelevant parameters they are uninformative. The mean of l_i and v are approximately equal for large α. The width of the peak is controlled by $\alpha - 2$ and $\alpha(D-d) - 2$ respectively, and $var(v) \approx \frac{1}{D-d} var(l_i)$. The fact that they have the same mode is consistent with our prior knowledge about the variance of both retained and discarded directions: Since l_i can be either larger or smaller than v, we don't place prior preference on this choice. The variance relationship between v and l_i is also consistent with the fact that at the maximum likelihood solution, v is the mean of $D-d$ discarded eigenvalues.

3.2 The MAP Estimator:

The full Bayesian approach is computationally expensive. In this paper, we use MAP estimation for the parameters, which is very fast and will be shown to perform well with a proper choice of hyper-parameters.

Multiplying the likelihood with the prior probability gives the posterior probability

$$p(U,L,v,m|\mathcal{D},\alpha) \propto |C^{-1}|^{n/2} \exp(-\frac{1}{2} tr(C^{-1}(S+\alpha\beta I))) \quad (7)$$

where $n = N + \alpha$. It's easy to show that the maximum of the posterior distribution is given by,

$$\hat{m} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad (8)$$

Plugging in $det(C^{-1}) = det(WW^T)v^{-(D-d)}$ and comparing the logarithm of the posterior distribution, equ 7 with equ 9 in [Welling *et al.*, 2003],

$$\mathcal{L} = -\frac{ND}{2} log(2\pi) + \frac{N}{2} log\, det(WW^T) + \frac{N(D-d)}{2} log(\frac{1}{\sigma_0^2}) - \frac{N}{2} tr(C_{XCA}^{-1} S)$$

we can see that these two equations are equivalent if we replace N in [Welling *et al.*, 2003] by n, and S by $(S+\alpha\beta I)/n$. Thus, the MAP estimator for U, L and v is obtained following the same derivation as in [Welling *et al.*, 2003]:

$$\hat{l}_i = \frac{N\lambda_i + \alpha\beta}{N+\alpha}, \quad i \in \mathcal{C} \quad (9)$$

$$\hat{v} = \frac{N}{N+\alpha}\frac{\sum_{i\in\mathcal{G}}\lambda_i}{D-d} + \frac{\alpha\beta}{N+\alpha} \quad (10)$$

where λ_i is the eigenvalue of the sample covariance matrix

$$\hat{S}/N = \sum_n (x_n - \hat{m})(x_n - \hat{m})^T/N \quad (11)$$

belonging to the set of $\mathcal{C}$. The ith column of U is the eigenvector of $\hat{S}/N$ corresponding to λ_i, and $\mathcal{C}$ and $\mathcal{G}$ are sets of respectively retained and discarded directions. As shown in [Welling *et al.*, 2003], all the discarded eigenvalues are contiguous in the ordered eigenspectrum of the sample covariance matrix, and G is determined by comparing all the $D-d+1$ candidate sets and choosing the set with the maximal posterior probability or equivalently with the minimal value of the following term:

$$\mathcal{K} = \sum_{i\in\mathcal{C}} \log \hat{l}_i + (D-d)\log(\sum_{i\in\mathcal{G}} \hat{l}_i) \quad (12)$$

The complete algorithm is summarized below:

Bayesian XCA MAP Solution

1. Compute the sample mean $\hat{m}$ and covariance matrix $\hat{S}/N$ using eqn 8, 11.
2. Compute the eigenvalues of $\hat{S}/N$ and get their estimates, $\hat{l}_i$, using eqn 9
3. Find the optimal set $\mathcal{G}$ among $D-d+1$ candidates, $\{j \le i \le j+d-1\}_{j=1}^{D-d+1}$ with the minimal value of $\mathcal{K}$ in eqn 12
4. Compute the estimate $\hat{v}$ using eqn 10

In the experiments of this paper, we remove the mean and normalize the variance of each attribute as a preprocessing step. The non-biased estimate of the mean eigenvalues of the covariance matrix is then always 1. Since the hyper-parameter β acts as the best prior guess for the eigenvalues, we set it to 1 in this paper. The parameter α is chosen through cross validation. We can also apply this prior to PCA (MCA) to derive a Bayesian PCA (MCA) model by simply imposing a constraint $l_i > v$ ($l_i < v$) on Bayesian XCA. The only difference is that no comparisons between eigenvalue sets are necessary because under the constraint Bayesian PCA (MCA) can only choose PCs (MCs).

The estimators $\hat{l}_i$ and $\hat{v}$ satisfy some interesting properties. They are both a linear combination of the un-regularized estimator and a constant with a weight proportional to the number of data points N. When the dataset is large enough ($N \gg \alpha$), the effect of sampling noise can be neglected and

the estimator $\hat{l}_i \approx \lambda_i$, $\hat{v} \approx \frac{1}{D-d}\sum_{i\in\mathcal{G}}\lambda_i$, equivalent to the ML solution of XCA. When the dataset becomes smaller or α is larger, the constant term gets more weight. The eigenvalue estimator then trades bias for variance. Moreover, in terms of the ordered eigenvalues, as mentioned at the beginning of this section, the un-regularized estimator has a strong bias, which is possibly larger than the bias induced by the prior especially when $N < D$. Therefore, with properly chosen hyper-parameters, $\hat{l}_i$, $\hat{v}$ can reduce both the variance and bias of the ordered eigenvalue estimators.

3.3 Other Eigenvalue Estimators

Besides the MAP estimators, there are also other kinds of estimators for the eigenvalues of the covariance matrix. One is introduced by Lawley [Lawley, 1956] and cited in [Jackson, 2003]. It is able to correct the bias of ordered sample eigenvalues for Gaussian distributions up to $O(1/N)$. We have implemented this method but did not include it in our experimental results because it was quite unstable and even produced negative variance estimates. Another estimator is a shrinkage method with an automatic selection of the shrinkage parameter [Schäfer and Strimmer, 2005][3]. It's designed for the case where the number of data instances is close to or even smaller than the number of attributes. This estimator was included in our experiments described below.

4 Mixtures of Bayesian XCA

Mixtures of factor analyzers [Ghahramani and Beal, 2000] and mixtures of PPCAs [Tipping and Bishop, 1999a; Bishop and Tipping, 1998] are potentially powerful density estimators that combine a number of local dimensionally reduced models into a single joint model. It is not unreasonable to assume that these local models represent constraints, i.e. local patches of data shaped like high dimensional pancakes. Unfortunately, the issue of overfitting is exacerbated for local mixtures because the effective number of data items available to estimate each model component is much smaller. We propose that a mixture of Bayesian XCA models might resolve this.

We therefore introduce priors for each mixture component with shared hyper-parameters. The objective function we want to maximize is the log-likelihood including the regularizing prior given by

$$L(\mathcal{D},\pi,\theta) = \log\left[\prod_n\left(\sum_{z_n} P_{XCA}(x_n|\theta_{z_n})\pi_{z_n}\right)\prod_k p(\theta_k)\right] \tag{13}$$

where $\theta = \{W, v, m\}$, z_n is the cluster index of x_n, and $\pi_k = P(z_n = k)$.

The EM algorithm is used to train this model. It has a regular E-step and in the M-step, we execute the BXCA algorithm for each mixture component. Iterating E- and M-steps until convergence is guaranteed to converge to a local maximum of the MAP regularized log-likelihood.

[3]The code is available at *http://strimmerlab.org/software.html*

5 Experiments

5.1 Log-likelihood:

In this section we will compare the test log-likelihood on various datasets for five different models: XCA, PCA, Bayesian XCA, Bayesian PCA, and Shrinkage XCA.

Mosquito Wing Landmarks:

In section 1, we have shown that the first PC and MC of the mosquito wing landmark data represent respectively the large variability of the locations of veins and a constraint on the shape of wings. In this experiment, we want to compare the log-likelihood for different models on this dataset. A number of landmarks on the boundary of the wing are sub-sampled, and the distances of each landmark to its 4 nearest neighbors along the edge are used as features. For the 8 landmarks, there are a total of 13 distances as shown in the middle of two subfigures in Figure 1(b). This is the same number of degrees of freedom after removing translation and rotation information from the original coordinates.

Figure 3(a) shows the average log-likelihood of data points in the training (solid) and test (dashed) sets for Bayesian XCA, XCA and PCA with different numbers of retained directions. We use 50 data points in the training set and 77 in the test set. The inset shows the number of MCs picked by Bayesian XCA and XCA. The log-likelihood of Bayesian XCA and XCA are everywhere above PCA. Bayesian XCA and XCA always pick MCs when $d < 11$, suggesting that it's better to model the data with constraints. Overfitting is not very serious in this experiment and the plots of Bayesian XCA and XCA are close to each other. Figure 3(b) shows the comparison of Bayesian XCA, Bayesian PCA and Shrinkage XCA w.r.t. the average log-likelihood of test data. Clearly Bayesian XCA performs much better than the other two.

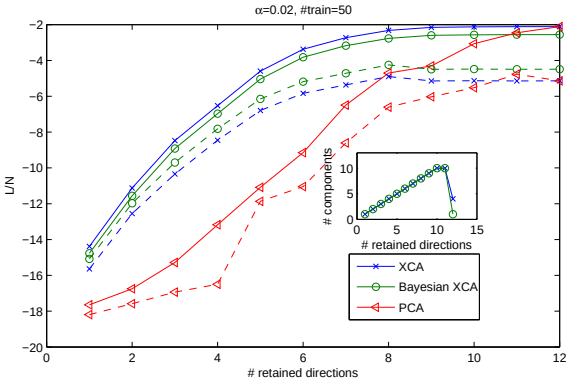
"Frey Faces" Image Data

We have repeated the experiment in [Welling *et al.*, 2003] on the "Frey Faces" images[4]. This dataset contains 1965 images of size 20×28. On each pixel, the values are normalized across images to zero mean and unit variance. Figure 4(a),(b) show plots of the average log-likelihood per data point for the various methods. 1000 samples are used in the training set and the remaining 965 in the test set. The corresponding numbers of minor components picked by Bayesian XCA and XCA are plotted in the inset.

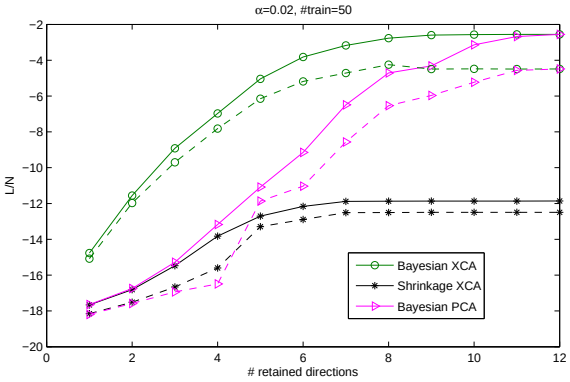
XCA and PCA are compared on the same dataset in [Welling *et al.*, 2003] and it turns out that XCA overfits quickly because of the small size of the training set (see Figure 4(a)). With the prior, Bayesian XCA doesn't suffer from this problem. In fact, it remains to perform robustly even when we retain a large number of dimensions. In effect, it will not pick minor components unless sufficiently supported by the data. As a result it outperforms both XCA and PCA. From Figure 4(b) we find that the performance of Bayesian XCA is close to that of Bayesian PCA and Shrinkage XCA underscoring the fact that overfitting is the real issue here.

Figure 5 show the log-likelihood and corresponding numbers of MCs on different sizes of training set. Unlike XCA,

[4]Obtained from *http://www.cs.toronto.edu/~roweis/data.html*

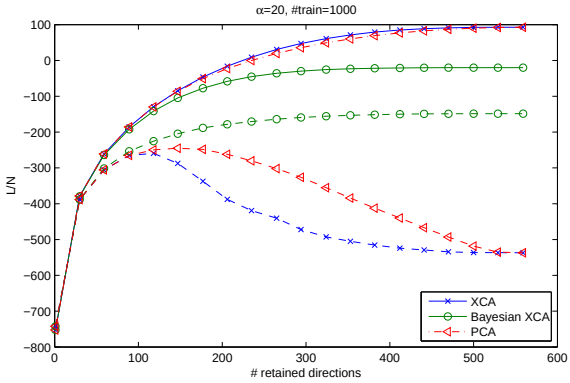


(a)

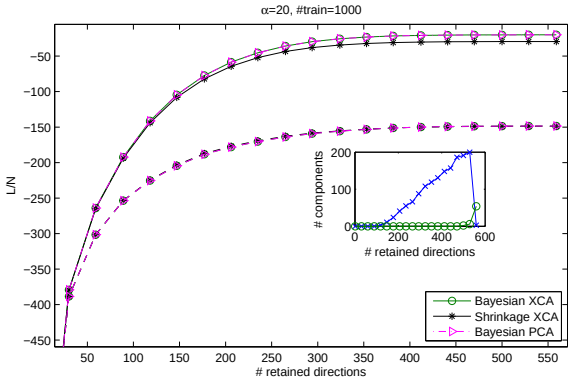


(b)

Figure 3: (a),(b) show the average log-likelihood per data point of the Mosquito Wing Landmark data on the training set of 50 data points (solid) and test set of 77 data points (dashed) as a function of the number of retained directions. (a) compares Bayesian XCA with XCA and PCA, and (b) compares Bayesian XCA with Bayesian PCA and Shrinkage XCA. Hyper-parameter $\alpha = 0.02$ is determined through cross validation with $d = 5$. The inset shows the number of minor components for different numbers of retained directions for Bayesian XCA (o) and XCA (x).



(a)



(b)

Figure 4: (a),(b) show the average log-likelihood per data point of the "Frey Faces" data on 1000 training data (solid) and 965 test data (dashed) as a function of the number of retained components. (a) compares Bayesian XCA with XCA and PCA, and (b) compares Bayesian XCA with Bayesian PCA and Shrinkage XCA. Hyper-parameter $\alpha = 20$ is determined through cross validation with $d = 300$. The inset shows the number of minor components for Bayesian XCA (o) and XCA (x).

Bayesian XCA always chooses principal components on this dataset, and thus exhibits similar performance as Bayesian PCA. While XCA and PCA over-fit severely as the number of data points decreases the two Bayesian models show good performance across the board on the test set.

In all experiments we determined the value for α using cross validation using a single value of d (nr. of retained dimensions) and N (training size) and subsequently used this hyper-parameter setting for all the other values of d and N.

The experiments suggest that the performance of Bayesian XCA is not very sensitive to the choice of α. To confirm that we also estimated the optimal value of α for a range of d and N values. We found that α was relatively stable and that the test log-likelihood only marginally improved. Results are not presented due to space limitation.

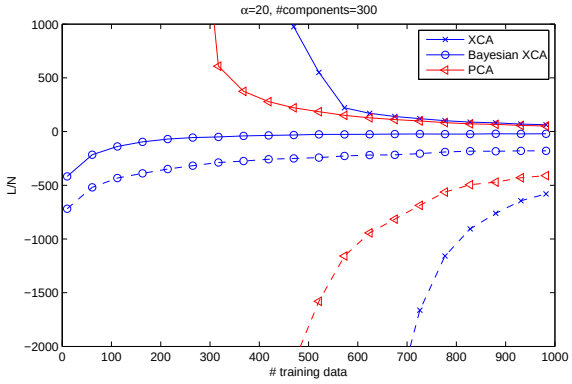
5.2 Mixture Models

We have also tested the performance of Bayesian XCA on a dataset from the UCSD data-mining competition[5] which has 19 attributes. These results are not included in this paper due to the space limitations. However, Bayesian XCA also shows its advantage to other methods on this dataset. Unlike the previous two experiments where either PCs or MCs were chosen by Bayesian XCA, on this dataset it finds a more balanced mix of the two.

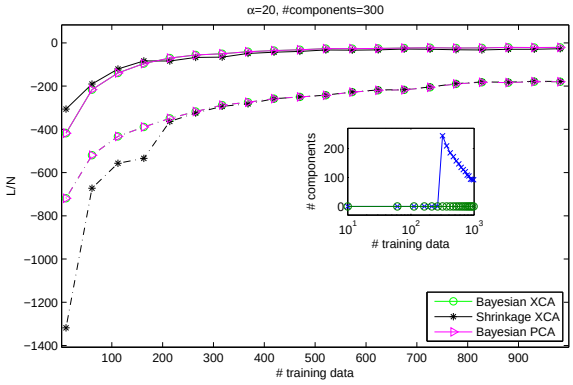
We have compared the performance of a mixture of Bayesian XCA models with mixtures of respectively Bayesian PCA, XCA and PCA. We have tested these models on the UCSD dataset. We ran the proposed EM algorithm on 1000 samples with another 1000 samples for validation and a further 38000 samples as the test set. For each model, there are 20 mixture components and the number of retained directions is 10. The EM algorithm was terminated in two ways: it either ran until convergence on training set, or was stopped early by monitoring performance on the validation set. Results averaged over 100 runs are shown in Table 1.

The best log-likelihood is obtained by Bayesian XCA with early stopping, and the worst by XCA running until EM converges. Generally, Bayesian models perform better than non-Bayesian models, and early stopping is better than running until convergence. However, the difference between these

[5]Obtained from *http://mill.ucsd.edu/index.php?page=Datasets &subpage=Download* in the standard classification task.



(a)



(b)

Figure 5: (a),(b) show the average log-likelihood per data point of the "Frey Faces" data on the training data (solid) and test data (dashed) as a function of the number of training data points. (a) compares Bayesian XCA with XCA and PCA, and (b) compares Bayesian XCA with Bayesian PCA and Shrinkage XCA. Hyperparameter $\alpha = 20$. The inset shows the number of minor components for Bayesian XCA (o) and XCA (x).

two stopping criteria for Bayesian models is much less pronounced than for non-Bayesian methods. This suggests that Bayesian methods are able to reduce overfitting to a substantial extent. In a mixture model, Bayesian XCA is still able to combine PCs and MCs reliably and gives a better estimate of the probability density than Bayesian PCA/MCA.

6 Conclusions

We have argued that oftentimes constraints (minor components) better characterize the structure of data than the widely adopted principal components. In particular, scientific discovery seems to be mainly based on discovering conservation laws which directly correspond to minor components.

The primary goal of this paper is to introduce a practical method to reliably extract the optimal combination of principal and minor components from data. An earlier method, called "extreme components analysis" (XCA) [Welling *et al.*, 2003] provided a first step towards this goal but was highly sensitive to sampling noise and as a result suffered from overfitting. The Bayesian extension of XCA we propose in this paper largely resolves this problem. This was verified empirically on various datasets. Moreover, Bayesian XCA provides a better density estimator than a number of alternative methods such PCA, Bayesian PCA, and XCA based on frequentist corrections to the spectrum. The proposed method is also highly scalable since it is based on a simple singular eigenvalue decomposition of the data matrix.

	Bayesian		Non-Bayesian	
	Covergence	Early Stop	Convergence	Early Stop
MCA	-18.65	-18.44	-22.57	-18.6
XCA	-18.51	**-18.31**	**-22.91**	-18.56
PCA	-19.44	-19.25	-22.4	-19.92

Table 1: Average log-likelihood of test set on UCSD dataset for mixtures of MCA, XCA and PCA with and without the prior. The EM algorithm stops when it is converged or when it is terminated by monitoring performance on a validation set. The maximal and minimal values are bold-faced.

Code for Bayesian XCA will be released to the public soon. We hope that this will facilitate its widespread use in the scientific community.

References

[Bishop and Tipping, 1998] Christopher M. Bishop and Michael E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):281–293, 1998.

[Bishop, 1999] Chris Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

[Ghahramani and Beal, 2000] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *NIPS*, volume 12, 2000.

[Jackson, 2003] Edward J. Jackson. *A User's Guide to Principal Components (Wiley Series in Probability and Statistics)*. Wiley-Interscience, September 2003.

[Lawley, 1956] D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1-2):128–136, 1956.

[Minka, 2000] Thomas P. Minka. Automatic choice of dimensionality for PCA. Technical report, M.I.T. Media Laboratory Perceptual Computing Section, 2000.

[Roweis, 1998] Sam Roweis. EM algorithms for PCA and SPCA. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 626–632, Cambridge, MA, USA, 1998. MIT Press.

[Schäfer and Strimmer, 2005] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 4, 2005.

[Tipping and Bishop, 1999a] M. E. Tipping and C. M. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2):443482, 1999.

[Tipping and Bishop, 1999b] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1999.

[Welling *et al.*, 2003] M. Welling, F. Agakov, and C.K.I. Williams. Extreme components analysis. In *Neural Information Processing Systems*, volume 16, Vancouver, Canada, 2003.

[Williams and Agakov, 2002] C.K.I. Williams and F.V. Agakov. Products of Gaussians and probabilistic minor components analysis. *Neural Computation*, 14(5):1169–1182, 2002.