CLASSIFYING NON-GAUSSIAN AND MIXED DATA SETS IN THEIR NATURAL PARAMETER SPACE

Cécile Levasseur[†], *Uwe F. Mayer*[‡] *and Ken Kreutz-Delgado*[†]

[†]Jacobs School of Engineering
University of California, San Diego
La Jolla, CA, USA
{*clevasseur, kreutz*}*@ucsd.edu*

[‡]Department of Mathematics
University of Utah
Salt Lake City, UT, USA
mayer@math.utah.edu

ABSTRACT

We consider the problem of both supervised and unsupervised classification for multidimensional data that are non-gaussian and of mixed types (continuous and/or discrete). An important subclass of graphical model techniques called Generalized Linear Statistics (GLS) is used to capture the underlying statistical structure of these complex data. GLS exploits the properties of exponential family distributions, which are assumed to describe the data components, and constrains latent variables to a lower dimensional parameter subspace. Based on the latent variable information, classification is performed in the natural parameter subspace with classical statistical techniques. The benefits of decision making in parameter space is illustrated with examples of categorical data text categorization and mixed-type data classification. As a text document preprocessing tool, an extension from binary to categorical data of the conditional mutual information maximization based feature selection algorithm is presented.

1. INTRODUCTION

The complexity of data generally comes from the possible existence of an extremely large number of components and from the fact that the components are often of mixed types, i.e., some components might be continuous (with different underlying distributions) and some components might be discrete (categorical, count or Boolean). This is typically the case in drug discovery, health care, or fraud detection.

Graphical models, also referred to as Bayesian Networks when their graph is directed, are a powerful tool to encode and exploit the underlying statistical structure of complex data sets [5]. The Generalized Linear Statistics (GLS) framework represents a subclass of graphical model techniques and includes as special cases multivariate probabilistic systems such as Principal Component Analysis (PCA), Generalized Linear Models (GLMs) and factor analysis [7]. It is equivalent to a computationally tractable mixed exponential families data-type hierarchical Bayes graphical model with latent variables constrained to a low-dimensional parameter subspace. The use of exponential family distributions allows the data components to have different parametric forms and exploits the division between data space and parameter space specific to exponential families. In addition to giving a generative model that can be fit to the data, it offers the advantage that problems can be attacked in a latent variable parameter subspace that is a continuous, Euclidean space, even when data are categorical or of mixed types.

Although a variety of techniques exists for performing inference on graphical models, it is, in general, very difficult to learn the parameters which constitute the model, even if it is assumed that the graph structure is known. The main goal of this paper is to demonstrate our ability to learn a generative GLS graphical model that captures the statistical structure of the data, to then use this knowledge to gain insight into the problem domain, and perform effective classification. The text categorization and classification problems shown in the paper serve this purpose as examples illustrating the benefits of making decisions in parameter space rather than in data space as done with more classical approaches. Support Vector Machines as well make decisions in a non-data space. However, although often promising the highest accuracy, this technique will not generally provide any better understanding of the data. An advantage of learning a generative model of the data as done with GLS is that generating synthetic data for the purposes of developing and training classifiers with the same statistical structure as the original data becomes possible. This is particularly useful in cases where data are very difficult or expensive to obtain, and when the original data are proprietary and cannot be directly used for publication purposes in open literature.

In this paper, we first review the GLS framework and show how natural it is for non-gaussian data of mixed types. Then we demonstrate the utility of this approach with experiments on real data sets, where classification in parameter space outperforms classification in data space.

This research was partially supported by NSF grant No. CCF-0830612.

978-1-4244-4948-4/09/$25.00 © 2009 IEEE

2. GENERALIZED LINEAR STATISTICS

The Generalized Linear Statistics framework is based on the hierarchical Bayes graphical model for hidden or latent variables shown in Figure 1 [7].

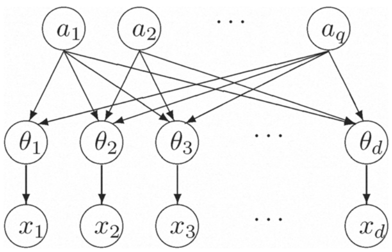

Fig. 1. Graphical model for the GLS framework.

The row vector $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ consists of observed features of mixed data instances in a d-dimensional space. It is assumed that instances can be drawn from populations having class-conditional probability density functions

$$p(\mathbf{x}|\boldsymbol{\theta}) = p_1(x_1|\theta_1) \cdot \ldots \cdot p_d(x_d|\theta_d), \tag{1}$$

where, when conditioned on the random parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d] \in \mathbb{R}^d$, the components of $\mathbf{x}$ are independent. The subscript i on $p_i(\cdot|\cdot)$ serves to indicate that the marginal densities can all be different, allowing for the possibility of $\mathbf{x}$ containing categorical, discrete, and continuous valued components. Also, the marginal densities are each assumed to be one-parameter exponential family densities, and θ_i is taken to be the natural parameter (or some simple bijective function of it) of the exponential family density p_i. Each component density $p_i(x_i|\theta_i)$ in (1) for $x_i \in \mathcal{X}_i$, $i = 1, \ldots, d$, is of the form

$$p(x_i|\theta_i) = \exp\big(\theta_i x_i - G(\theta_i)\big),$$

where $G(\cdot)$ is the cumulant generating function defined as

$$G(\theta_i) = \log \int_{\mathcal{X}_i} \exp\big(\theta_i x_i\big) \nu(\mathrm{d}x_i),$$

with $\nu(\cdot)$ a σ-finite measure that generates the exponential family. It can be shown that $G(\boldsymbol{\theta}) = \sum_{i=1}^{d} G(\theta_i)$ [7].

It is further assumed that $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\theta} = \mathbf{a}\mathbf{V} + \mathbf{b} \tag{2}$$

with the hidden or latent variable $\mathbf{a} = [a_1, \ldots, a_q] \in \mathbb{R}^q$ random and unknown with $q < d$ (and ideally $q \ll d$), $\mathbf{V} \in \mathbb{R}^{q \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ deterministic and unknown. The latent variable $\mathbf{a}$ in some way explains part (or all) of the random behavior of the observed variables.

The maximum likelihood identification of the blind random effect model

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int \prod_{i=1}^{d} p_i(x_i|\theta_i)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{3}$$

with $\pi(\boldsymbol{\theta})$ the probability density function of $\boldsymbol{\theta}$, is quite a difficult problem. It corresponds to identifying $\pi(\boldsymbol{\theta})$, which, under the condition $\boldsymbol{\theta} = \mathbf{a}\mathbf{V} + \mathbf{b}$, corresponds to identifying the matrix $\mathbf{V}$, the vector $\mathbf{b}$, and a density function on the random effect $\mathbf{a}$ via a maximization of the likelihood function $p(\mathbf{X})$ with respect to $\mathbf{V}$, $\mathbf{b}$, and the random effect density function, where

$$p(\mathbf{X}) = \prod_{k=1}^{n} p(\mathbf{x}[k]) = \prod_{k=1}^{n} \int p\big(\mathbf{x}[k]|\boldsymbol{\theta}\big)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4}$$

and $\mathbf{X}$ is the $(n \times d)$ observation matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}[1] \\ \mathbf{x}[2] \\ \vdots \\ \mathbf{x}[n] \end{pmatrix} = \begin{pmatrix} x_1[1] & \ldots & x_d[1] \\ x_1[2] & \ldots & x_d[2] \\ \vdots & \ddots & \vdots \\ x_1[n] & \ldots & x_d[n] \end{pmatrix}.$$

This difficulty can be avoided by Non-Parametric Maximum Likelihood (NPML) estimation of the random effect distribution, concurrently with the structural model parameters. The NPML estimate is known to be a discrete distribution on a finite number of support points [6, 8]. As shown in [7], the NPML approach yields unknown point-mass support points $\underline{\mathbf{a}}[l]$, point-mass probability estimates π_l, and the linear predictor $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}$ for $l = 1, \ldots, m, m \leq n$. The single-sample likelihood (3) then becomes

$$p(\mathbf{x}) = \sum_{l=1}^{m} p\big(\mathbf{x}|\underline{\boldsymbol{\theta}}[l]\big)\pi_l = \sum_{l=1}^{m} p\big(\mathbf{x}|\underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}\big)\pi_l$$

and the data likelihood (4) is equal to

$$p(\mathbf{X}) = \prod_{k=1}^{n}\sum_{l=1}^{m} p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big)\pi_l = \prod_{k=1}^{n}\sum_{l=1}^{m} p\big(\mathbf{x}[k]|\underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}\big)\pi_l.$$

The data likelihood is thus approximately the likelihood of a finite mixture of exponential family densities with unknown mixture proportions or point-mass probability estimates π_l and unknown point-mass support points $\underline{\mathbf{a}}[l]$, with the linear predictor $\underline{\boldsymbol{\theta}}[l]$ in the lth mixture component. The combined problem of maximum likelihood estimation of the parameters $\mathbf{V}$, $\mathbf{b}$, the point-mass support points $\underline{\mathbf{a}}[l]$ and the point-mass probability estimates $\pi_l, l = 1, \ldots, m$, can be attacked by either using the Expectation-Maximization algorithm [3, 6, 8, 1], as done in particular in the Semi-Parametric Principal Component Analysis technique [11], or by simply considering the special case of uniform point-mass probabilities, i.e., $\pi_l = 1/m \ \forall l$, for which the number of support points equals the number of data samples. It was demonstrated in [7] that this special uniform case corresponds to the exponential Principal Component Analysis technique [2]. We are using this special case in this paper.

3. CLASSIFYING IN PARAMETER SPACE: REAL DATA EXPERIMENTAL RESULTS

The data sets used in this work are from the UC Irvine machine learning repository [14]. For each data set we do as follows. For text categorization examples, data preprocessing is needed, including a dictionary learning step. Then, for each data set, a low-dimensional latent variable subspace is identified in parameter space using GLS. In data space, classical Principal Component Analysis selects a lower dimensional subspace. Finally, classification is performed on both subspaces and performances are compared.

3.1. Text Categorization

The Twenty Newsgroups and the Reuters-21578 data sets account for most of the experimental work in text categorization, one example of information retrieval tasks. Text categorization is the activity of labeling natural language texts with thematic categories from a predefined set [13].

It has been acknowledged by the text categorization community that words seem to work well as features of a document for many classification tasks. In addition, it is usually assumed that the ordering of the words in a document does not matter. Hence, a document can be represented as a vector for which each distinct word is a feature [9]. There are two ways to characterize the value of each feature that are commonly used in the literature: Boolean and *tf*×*idf* weighting schemes. In Boolean weighting, the weight of a word is 1 if the word appears in the document and 0 otherwise. We choose to characterize the value of each feature by using the *tf*×*idf* scheme as recently more commonly used for document representation [12, 13]. The *tf*×*idf* weight is a statistical measure used to evaluate how important a word (or term) is to a corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The *term frequency tf* is the number of times a specific word occurs in a specific document. The *document frequency df* is the number of documents in which the specific word occurs at least once. The *inverse document frequency idf* is calculated from the document frequency, yielding the *tf*×*idf* weight w_i for feature i:

$$w_i = tf_i \cdot idf_i = tf_i \cdot \log\left((\text{total \# of documents})/df_i\right).$$

3.1.1. *Twenty Newsgroups data set*

The Twenty Newsgroups data set consists of Usenet articles collected from twenty different newsgroups. Each newsgroup contains 1000 articles. We consider the three following newsgroups: sci.med, comp.sys.mac.hardware and comp.sys.ibm.pc.hardware. We decide on a text categorization problem with two distinct classes, the first class consisting of the newsgroup sci.med and the second class consisting of the two other newsgroups.

Following the text document representation preprocessing steps described in Figure 2, we first choose to discard all header fields such as Cc, Bcc, Message-ID, as well as the Subject field (this step is called parsing). Case-folding is performed by converting all the characters into lower-case. We use a stop list, i.e., a list of words that will not be taken into account. Indeed, there are words such as pronouns, prepositions and conjunctions which are encountered very frequently but carry no useful information about the content of the document. We used a stop list commonly used in the literature, ftp://ftp.cs.cornell.edu/pub/smart/english.stop. It consists of 571 stop-words and yields a drastic reduction in the number of features. Then, some simple stemming is performed, such as removing the third person and plural "s". In addition to removing very frequent words with the stop list, we remove words appearing less than 10 times in the corpus. The *tf*×*idf* weighting scheme is then used and we choose to bin the weights and work with integer valued weights (5 bins are selected), i.e., categorical features.

Modified dictionary learning: Last, we construct a dictionary, and hence reduce the dimensionality of the feature space. There are various methods commonly applied for dimensionality reduction in document categorization [9]. We choose a conditional mutual information based approach to select a dictionary of d = 150 words. We modify the binary feature selection with conditional mutual information algorithm proposed in [4] to fit a categorical feature. The feature selection algorithm proposed in [4] is based on the conditional mutual information maximization criterion and selects features that maximize both the information about the class and the independence between features. The modification from binary to categorical is simple: following the definition of entropy and mutual information shown in [4], the summations are changed from summing over two values to summing over the total number of bins values.

We use this data set leaving out a randomly selected 40% of the instances of each class to use as a test set. The training set then consists of 1764 instances and the test set 1236. The dictionary is learned using the training set only.

Classification effectiveness is often measured in terms of *precision* and *recall* in the text categorization community [13]. Precision with respect to a class $\mathcal{C}_i$ (π_i) is defined as the probability that, if a random document is classified under $\mathcal{C}_i$, this decision is correct. Recall with respect to a class $\mathcal{C}_i$ (ρ_i) is defined as the probability that, if a random document ought to be classified under $\mathcal{C}_i$, this decision is taken. These probabilities are estimated in terms of the contingency table for $\mathcal{C}_i$ on a given test set as follows:

$$\widehat{\pi}_i = \frac{TP_i}{TP_i + FP_i} \quad \text{and} \quad \widehat{\rho}_i = \frac{TP_i}{TP_i + FN_i},$$

where TP_i, FP_i and FN_i refer to the sets of *true positives*

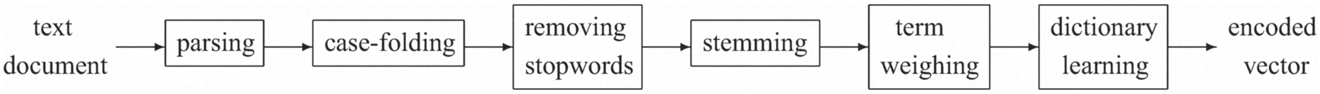

Fig. 2. Preprocessing and document representation for text categorization.

with respect to $\mathcal{C}_i$ (documents correctly deemed to belong to class $\mathcal{C}_i$), *false positives with respect to* $\mathcal{C}_i$ (documents incorrectly deemed to belong to class $\mathcal{C}_i$), and *false negatives with respect to* $\mathcal{C}_i$ (documents incorrectly deemed not to belong to class $\mathcal{C}_i$). Then, the F_1 measure combines precision and recall, attributing equal importance to π and ρ:

$$F_1 = \frac{2 \cdot \pi\rho}{\pi + \rho}.$$

When effectiveness is computed for several classes, the results for individual classes can be averaged in two ways: *microaveraging*, where π and ρ are obtained by summing over all individual classes (the subscript "μ" indicates microaveraging), and *macroaveraging*, where π and ρ are first evaluated "locally" for each class and then "globally" by averaging over the results of the different classes (the subscript "M" indicates macroaveraging) [13].

Supervised text categorization: Table 1 compares classification performances on (a) the q-dimensional latent variable subspace learned with GLS using a Binomial distribution assumption and (b) the q-dimensional classical PCA subspace learned in data space in terms of precision, recall and F_1 measure, for several values of q. The classifier is a simple linear discriminant. The classification performances are often very similar, at times at the advantage of GLS (q = 4 and 10), at other times at the advantage of classical PCA.

Unsupervised text categorization: The K-means algorithm is used to cluster the training documents into two distinct classes. Based on this clustering information, a linear discriminant is learned on the training documents and used to classify the test documents. Figure 3 presents the corresponding ROC curve for this unsupervised approach performed on both the GLS parameter subspace and the classical PCA data subspace (q = 2). The performance is best when the unsupervised approach is used on the GLS subspace rather than on the classical PCA subspace. In this example, even though it is of interest, we did not further investigate the impact of the value for q on the performance.

3.1.2. Reuters-21578 data set

The Reuters-21578 text categorization test collection Distribution 1.0 is considered as the standard benchmark for automatic document organization systems and consists of documents that appeared on the Reuters newswire in 1987. This corpus contains 21578 documents assigned to 135 different economic subject categories called *topics*. The topics are not disjoint. For the training test division of the data, the "Modified Apte" (ModApte) split is used. We reduce the size of the training test sets by only considering the ten topics that have the highest number of training documents as commonly done in the literature [13]. These topics are given in Table 2 and yield a training set of 6490 documents and a test set of 2545 documents. They cover almost all of the data, hence, researchers are able to restrict their work to them and still capture the essence of the data set.

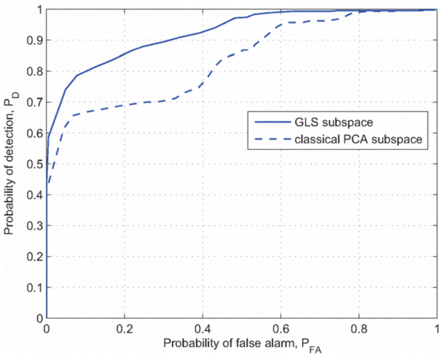

Fig. 3. Twenty Newsgroups data set: ROC curve for the proposed unsupervised text categorization technique performed on the low-dimensional subspace learned by (a) GLS (solid line) and (b) classical PCA (dashed line) ($q = 2$).

The data are preprocessed as done for the previous data set: parsing, case-folding, elimination of stopwords, stemming by using Porter's stemming algorithm commonly used for word stemming in English [10], elimination of words that appear less than 20 times in the corpus, *tf*×*idf* weighting. Then, we choose to bin the weights and work with integer valued weights (5 bins are selected), i.e., categorical features. A dictionary of d = 50 words is learned using the following approach. The dictionary is learned on the training set only and built independently for each of the ten classes. Feature selection was incremental. First we do a backward selection to 300 features with linear regression. From these 300 features, we use a logistic regression with a number of iterations reduced down to 5 for convergence, and do a backward selection down to 100 features. Finally, we do a standard full-convergence logistic regression from

Table 2. The ten topics with the highest number of training documents in the Reuters-21578 data set with the number of their documents in the training and test sets.

topics	training set	test set
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	369	118
interest	347	131
wheat	212	71
ship	197	89
corn	181	56

those 100 features down to 50 features.

Table 3 compares classification performances micro- and macroaveraged over the top ten categories of the Reuters-21578 data set using a linear discriminant classifier on (a) the latent q-dimensional variable subspace learned with GLS using a Binomial distribution assumption and (b) the classical PCA q-dimensional subspace learned in data space. Microaveraging and macroaveraging methods give quite different results: the linear discriminant classifier performs better based on the GLS information than on classical PCA information when the macroaveraging method is used, while microaveraging emphasizes how similar the two results are. It is known that the ability of a classifier to behave well on categories with few positive training instances will be highlighted by macroaveraging compared to microaveraging [13]. The linear discriminant classifier based on GLS information performs very well for the categories with fewer positive training instances yielding a better macroaveraged performance than the microaveraged one, cf. Table 2.

3.2. Abalone data set

The task is to predict the age of an abalone based on physical measurements. The Abalone data set consists of 4177 instances with 8 attributes. The problem can be seen as a classification problem aiming to distinguish three classes (number of rings = 1 to 8, number of rings = 9 to 10, number of rings = 11 and higher). We use this data set leaving out a randomly selected 40% of the instances to use as a test set (2506 training points and 1671 test points). Attribute 1 (sex, defined as infant, male or female) is the only noncontinuous attribute. We choose to model this attribute with a Binomial distribution, hence choosing a Gaussian-Binomial mixed-data assumption. Table 4 compares micro- and macroaveraged classification performances using a linear discriminant classifier on (a) the latent q-dimensional variable subspace learned with GLS using a mixed Gaussian-Binomial distribution assumption and (b) the classical PCA q-dimensional subspace learned in data space. Performances are best when classification is performed on the GLS parameter subspace.

4. CONCLUSION

As with Bayesian Networks in general, the strength of the Generalized Linear Statistics framework is that it offers important insight into the underlying statistical structure of complex data, both creating a generative model of the data and making effective classification decisions possible. The benefits of making decisions in parameter space rather than in data space as done with more classical approaches have been illustrated with examples of Binomial data supervised and unsupervised text categorization and Gaussian-Binomial mixed-data supervised classification. However, one noticeable weakness of the framework is its running time. In addition, for the text categorization situation, the conditional mutual information maximization based feature selection algorithm was modified to fit categorical data.

5. REFERENCES

[1] D. Boehning, *Computer-Assited Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping, and Others*, Chapman and Hall/CRC, 2000.

[2] M. Collins, S. Dasgupta and R. Shapire, "A generalization of principal component analysis to the exponential family," *Advances in Neural Information Processing Systems*, 2001.

[3] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Royal Statistical Soc., Series B*, vol. 39, no. 1, pp. 1-38, 1977.

[4] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Machine Learning Research*, vol. 5, pp. 1531-1555, 2004.

[5] M. I. Jordan and T. J. Sejnowski, *Graphical Models: Foundations of Neural Computation*, MIT Press, 2001.

[6] N. Laird, "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *J. American Statistical Assoc.*, vol. 73, no. 364, pp. 805-811, 1978.

[7] C. Levasseur, B. Burdge, K. Kreutz-Delgado and U. F. Mayer, "A unifying viewpoint of some clustering techniques using Bregman divergences and extensions to mixed data sets," *Proc. 1st IEEE Int'l Workshop on Data Mining and Artificial Intelligence*, pp. 56-63, 2008.

[8] B. G. Lindsay, "The geometry of mixture likelihoods: a general theory," *Annals of Statistics*, vol. 11, no. 1, pp. 86-94, 1983.

[9] A. Özgür, "Supervised and unsupervised machine learning techniques for text document categorization," master's thesis, *Computer Eng., Bogazici University, Istanbul, Turkey*, 2004.

[10] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[11] Sajama and A. Orlitsky, "Semi-parametric exponential family PCA," *Advances in Neural Information Processing Systems*, 2004.

[12] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[13] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[14] http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Table 1. Twenty Newsgroups data set: linear discriminant classification performances on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution (1236 test instances).

	PCA - Precision	PCA - Recall	PCA - F_1	GLS - Precision	GLS - Recall	GLS - F_1
$q = 1$	0.5045	0.8149	**0.6232**	0.3677	0.6603	0.4744
$q = 2$	0.7843	0.9351	**0.8531**	0.7844	0.8918	0.8346
$q = 3$	0.9388	0.8846	**0.9109**	0.8641	0.8558	0.8599
$q = 4$	0.9389	0.8870	0.9122	0.8830	0.9615	**0.9206**
$q = 5$	0.9038	0.9712	**0.9363**	0.8931	0.9639	0.9272
$q = 6$	0.9038	0.9712	**0.9363**	0.8914	0.9663	0.9273
$q = 8$	0.9040	0.9736	**0.9375**	0.8813	0.9639	0.9208
$q = 10$	0.8904	0.9760	0.9312	0.9691	0.9038	**0.9353**

Table 3. Reuters-21578 data set: linear discriminant classification performances (microaveraged in (a) and macroaveraged in (b)) on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution.

(a) Microaveraged performances

	PCA - Precision$^\mu$	PCA - Recall$^\mu$	PCA - F_1^μ	GLS - Precision$^\mu$	GLS - Recall$^\mu$	GLS - F_1^μ
$q = 1$	0.2408	0.7306	0.3622	0.2845	0.5653	**0.3785**
$q = 2$	0.3704	0.8303	0.5123	0.4087	0.7665	**0.5331**
$q = 3$	0.4553	0.8296	**0.5880**	0.4239	0.8099	0.5565
$q = 4$	0.4709	0.8260	**0.5998**	0.4743	0.8128	0.5990
$q = 5$	0.6178	0.8275	**0.7075**	0.6233	0.7895	0.6966
$q = 6$	0.6265	0.8364	0.7164	0.6484	0.8056	**0.7185**

(b) Macroaveraged performances

	PCA - Precision^M	PCA - Recall^M	PCA - F_1^M	GLS - Precision^M	GLS - Recall^M	GLS - F_1^M
$q = 1$	0.2200	0.6403	0.2905	0.3040	0.6274	**0.3751**
$q = 2$	0.3763	0.7717	0.4475	0.4006	0.7174	**0.4757**
$q = 3$	0.4342	0.7842	0.5184	0.4662	0.7552	**0.5267**
$q = 4$	0.4594	0.7820	0.5423	0.5138	0.7611	**0.5804**
$q = 5$	0.4988	0.7673	0.5870	0.5307	0.7386	**0.6007**
$q = 6$	0.5306	0.7809	**0.6150**	0.5471	0.7373	0.6134

Table 4. Abalone data set: linear discriminant classification performances (microaveraged in (a) and macroaveraged in (b)) on the q-dimensional latent variable space learned with classical PCA and GLS with a Gaussian-Binomial mixed distribution.

(a) Microaveraged performances

	PCA - Precision$^\mu$	PCA - Recall$^\mu$	PCA - F_1^μ	GLS - Precision$^\mu$	GLS - Recall$^\mu$	GLS - F_1^μ
$q = 1$	0.5036	0.7120	0.5899	0.5043	0.7409	**0.6001**
$q = 2$	0.5085	0.7337	0.6007	0.5178	0.7385	**0.6088**

(b) Macroaveraged performances

	PCA - Precision^M	PCA - Recall^M	PCA - F_1^M	GLS - Precision^M	GLS - Recall^M	GLS - F_1^M
$q = 1$	0.5204	0.7126	0.5952	0.5208	0.7415	**0.6058**
$q = 2$	0.5242	0.7335	0.6062	0.5337	0.7380	**0.6141**