BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 225–232
doi:10.1093/bioinformatics/btu618

Gene expression

Advance Access publication September 29, 2014

GLAD: a mixed-membership model for heterogeneous tumor subtype classification

Hachem Saddiki[1,2,†], Jon McAuliffe[3,†] and Patrick Flaherty[1,4,*]

[1]Department of Biomedical Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA, [2]School of Science and Engineering, Al Akhawayn University, Ifrane, 53000, Morocco, [3]Department of Statistics, University of California, Berkeley, CA 94720, USA, and [4]Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA 01609, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Genomic analyses of many solid cancers have demonstrated extensive genetic heterogeneity between as well as within individual tumors. However, statistical methods for classifying tumors by subtype based on genomic biomarkers generally entail an all-or-none decision, which may be misleading for clinical samples containing a mixture of subtypes and/or normal cell contamination.

Results: We have developed a mixed-membership classification model, called GLAD, that simultaneously learns a sparse biomarker signature for each subtype as well as a distribution over subtypes for each sample. We demonstrate the accuracy of this model on simulated data, in-vitro mixture experiments, and clinical samples from the Cancer Genome Atlas (TCGA) project. We show that many TCGA samples are likely a mixture of multiple subtypes.

Availability: A python module implementing our algorithm is available from http://genomics.wpi.edu/glad/

Contact: pjflaherty@wpi.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 21, 2013; revised on August 4, 2014; accepted on September 12, 2014

1 INTRODUCTION

Genomic analyses of many solid cancers have demonstrated extensive genetic heterogeneity between (Bonavia *et al.*, 2011; Parsons *et al.*, 2008) as well as within individual tumors (Dexter *et al.*, 1978; Heppner, 1984). Microdissection and DNA sequencing have recently validated at high resolution the existence of intratumor heterogeneity (Gerlinger *et al.*, 2012). Characterizing such heterogeneity is important because it may be a contributing factor to mono therapy treatment failure (Gerlinger *et al.*, 2012). Accurately detecting subtypes in an individual tumor may lead to improved combinatorial therapies.

Many cancers have been classified into distinct genetic subtypes that develop by means of activation or repression of different driver pathways. These tumor subtypes are commonly identified and characterized by clustering the genomic data from hundreds of samples (Eisen *et al.*, 1998; Hofree *et al.*, 2013). In an effort to disambiguate driver from passenger mutations, the genomic signatures associated with each subtype are sparse and comprised only of those aberrations that are thought to be involved in oncogenesis. New tumors are then classified based on their similarity to the centroids or signatures of those subtypes. However, this classification approach makes an all-or-none assumption about the primary tumor that is incorrect for heterogeneous tumors.

Mixture models have been used extensively to analyze gene expression patterns in complex experiments. Gasch and Eisen (2002) used fuzzy k-means clustering to identify functionally co-regulated transciptional networks. Brunet *et al.* (2004) took a less heuristic non-negative matrix factorization (NMF) approach to decompose the gene expression data matrix into a product of a meta-gene matrix and a sample weight matrix. The NMF approach was extended to allow for sparseness in either of the factor matrices (Hoyer, 2004).

Mixed-membership models have emerged in recent years as a tool for data where the all-or-none clustering assumption is inappropriate. In text classification, the topic-modeling framework, which includes mixed-membership models, captures the structure in large document corpora, where each document may exhibit a mixture of topics (Blei *et al.*, 2003). Mixed-membership models have been used in population genetics (Falush *et al.*, 2007), social network analysis (Airoldi *et al.*, 2008), and elsewhere (Erosheva *et al.*, 2004; Wang and McCallum, 2006).

Our model achieves the dual purposes of (i) representing each sample as a mixture of genomic subtypes, and (ii) representing each subtype signature as a sparse set of genomic features that delineate driving oncogenic pathways. Our model provides a more general framework for representing mixed samples than all-or-none classification methods and we show that we obtain a more accurate estimate of mixture proportions than a mixed-membership model without subtype sparsity. We demonstrate our model on RNA expression data from primary glioblastomas (GBM), comprising thousands of genomic features and hundreds of samples. There we show that we recover known subtypes with a sparse set of driving aberrations, and we give evidence that many of the primary samples are mixed.

2 MODEL STRUCTURE

We are given a data matrix $y \in \mathbb{R}^{M \times N}$, where the element y_{ji} is an observation of feature j in sample i. We would like to

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

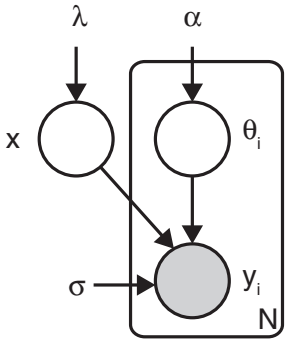

Fig. 1. Graphical model representation of the GLAD statistical model

represent each column of y as $y_i = x\theta_i$, where $x \in \mathbb{R}^{M \times K}$ is a matrix of K cluster centroids and $\theta_i \in \Delta^{K-1}$ is the ith sample's distribution over the K clusters. Furthermore, we would like x to be sparse, for purposes of cluster interpretability and generalizability to test cases. In the specific case of cancer subtyping, y_{ji} is a possibly normalized gene expression measurement for gene j in sample i.

2.1 Generative process

We introduce GLAD, a Gaussian–Laplace–Dirichlet model for mixed-membership data where the underlying clusters have a sparse representation. The name refers to the component distributions comprising the joint model. The generative process for the observed data involves unobserved cluster centroids and unobserved mixing proportions, as follows:

(1) Draw the elements of the $M \times K$ cluster-centroid matrix x as iid Laplace(λ) variables.

(2) For each sample $i = 1, \ldots, N$, independently:

(a) choose a distribution over clusters, $\theta_i \sim \text{Dirichlet}(\alpha)$;

(b) draw $y_i \mid \theta_i, x \sim \text{Normal}(x\theta_i, \Sigma)$.

This process involves several hyperparameters: λ, a non-negative scalar governing the Laplace prior on the elements of x; α, a K-vector of non-negative values controlling the possibly asymmetric Dirichlet prior on each sample's mixing proportions; and $\Sigma \in \mathbb{R}^{M \times M}$, the conditional covariance matrix for the normally distributed components of the gene expression vectors.

GLAD has two levels of latent-variable sampling: global and local. Globally, the cluster centroids are chosen once in advance for the entire dataset. Locally, the expected value of each observed vector is chosen as a mixture of cluster centroids, with different mixing weights for different observed vectors. Figure 1 depicts GLAD as as a graphical model.

The Laplace distribution over the x_{jk} enforces sparsity in the subtype matrix (Kabán, 2007). The automatic relevance determination method similarly uses a Laplace prior over features to induce sparsity (MacKay, 1992). The Laplace prior is also used to create sparsity in the multinomial inverse regression model for text sentiment analysis (Taddy, 2012).

As θ_i has a Dirichlet distribution, it provides, for each sample, a distribution over the K clusters. For simplicity, we set $\Sigma = \sigma^2 I_M$ from now on. The development of the model is similar for anisotropic or non-diagonal Σ, though some regularization may be required.

Under the GLAD model, the joint distribution of observed and latent variables for the ith observed sample, also called the complete likelihood, is as follows:

$$\begin{aligned} p(x, \theta_i, y_i; \alpha, \lambda, \sigma^2) &= p(x; \lambda)\, p(\theta_i; \alpha)\, p(y_i \mid x, \theta_i; \sigma^2) \\ &= \left[\prod_{j=1}^{M}\prod_{k=1}^{K} \frac{1}{2\lambda}\exp\left(-\frac{|x_{jk}|}{\lambda}\right)\right]\left[\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{ki}^{\alpha_k - 1}\right] \\ &\cdot \left[\prod_{j=1}^{M}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(y_{ji} - x_j\theta_i)^2\right)\right]. \end{aligned} \tag{1}$$

Integrating over the latent variables x and θ_i yields the marginal likelihood of a sample,

$$p(y_i; \alpha, \lambda, \sigma^2) = \int_x p(x; \lambda) \int_{\theta_i} p(\theta_i; \alpha)\, p(y_i \mid x, \theta_i; \sigma^2)\, d\theta_i\, dx. \tag{2}$$

Finally, under the independent-samples assumption, the likelihood of the full dataset is as follows:

$$p(y; \alpha, \lambda, \sigma^2) = \prod_{i=1}^{N} \int_x p(x; \lambda) \int_{\theta_i} p(\theta_i; \alpha)\, p(y_i \mid x, \theta_i; \sigma^2)\, d\theta_i dx. \tag{3}$$

The inference algorithms in Section 3 focus on the maximum likelihood estimate, that is, maximizing Equation (3) with respect to the hyperparameters α, λ, and σ^2.

2.2 Comparison with other clustering and topic models

GLAD differs from other gene expression clustering models. A classical clustering mixture model would sample a Dirichlet once for the entire dataset. The resulting probabilities would be used to assign each expression vector to one cluster. This is the assumption of the PAM50 method (Parker *et al.*, 2009) and other all-or-none clustering approaches (Koboldt *et al.*, 2012). Also, our method does not require side information, such as biochemical network structure, though this information may be incorporated into a structured sparsity parameter (Hofree *et al.*, 2013).

GLAD also differs from other topic models. A standard topic model chooses from a Dirichlet distribution once for each sample as we have done here. But then, given that Dirichlet draw, a number of additional per-sample latent variables are drawn, rather than just one as we have done in GLAD. Particular instances of this alternative structure are latent Dirichlet allocation (Blei *et al.*, 2003), where many additional multinomial random variables are drawn, and latent process decomposition (LPD) (Rogers *et al.*, 2005), where additional Gaussians are drawn. The sparseTM model for text classification also uses additional multinomial random variables, with another type of sparsifying prior (Wang and Blei, 2009).

2.3 Interpretation as matrix factorization

Many statistical models have corresponding matrix factorization interpretations (Singh and Gordon, 2008). Principal component analysis, NMF and latent semantic indexing have both linear-algebraic and probabilistic formulations. GLAD is different from standard decomposition approaches because of the constraint that θ_i form a distribution over the columns of x. In particular, the negative complete log likelihood, retaining only terms involving x and θ, is as follows:

$$-LL(x,\theta\,|\,y) = \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{1}{2\,\sigma^2}\left(y_{ji}-x_j\theta_i\right)^2 + N\sum_{j=1}^{M}\sum_{k=1}^{K}\frac{|\,x_{jk}\,|}{\lambda} - \sum_{k=1}^{K}(\alpha_k-1)\sum_{i=1}^{N}\log\theta_{ki}, \tag{4}$$

with the simplicial constraints $\theta_i^{\top}\mathbf{1}=1$ and $\theta_i \geq \mathbf{0}$ (elementwise) for each i. As we explain in Section 3, approximately minimizing Equation (4) is the main ingredient in the GLAD inference procedure.

To see the connection to matrix factorization, set $\sigma^2=\frac{1}{2}$, $\alpha=\mathbf{1}$, and $\lambda=1$, leaving the objective function as following:

$$f(x,\theta\,|\,y) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ji}-x_j\theta_i\right)^2 + N\sum_{j=1}^{M}\sum_{k=1}^{K}|x_{jk}\,|\,. \tag{5}$$

The first term in (5) can be viewed as a loss function for the difference between the fitted value $\hat{y}_i = x\theta_i$ and the actual observation y_i. The second term is a sparsifying regularizer, wherein setting x_{jk} to a non-zero value has a cost proportional to the magnitude of x_{jk}. Viewing Equation (5) as a Lagrangian function where only the terms involving x and $\theta=[\theta_1\ \ldots\ \theta_N]$ have been retained, the matrix factorization problem corresponding to GLAD can be written as follows:

$$\begin{aligned}&\text{minimize}_{x,\theta}\ ||\,y-x\theta\,||_2^2\\ &\text{subject to}\ ||\,x\,||_1 \leq b\\ &\theta^T\mathbf{1}_K=\mathbf{1}_N\\ &\theta\geq\mathbf{0}.\end{aligned} \tag{6}$$

The inequality on θ is element-wise. The optimization problem (6) is biconvex in x and θ. It bears some resemblance to sparse coding (Lee *et al.*, 2006), with interesting differences: in sparse coding, x is dense and θ is sparse and non-negative, whereas in (6), x is sparse and θ is dense and simplicial.

3 INFERENCE AND PARAMETER ESTIMATION

Inference in GLAD focuses on the posterior distribution over latent variables

$$p(\theta,x\,|\,y;\alpha,\lambda,\sigma^2) = \frac{p(\theta,x,y;\alpha,\lambda,\sigma^2)}{p(y;\alpha,\lambda,\sigma^2)}. \tag{7}$$

As exact inference is intractable, we have developed a non-conjugate variational inference algorithm (Wang and Blei, 2013) to estimate the posterior distribution $p(\theta,x\,|\,y;\hat{\alpha},\hat{\lambda},\hat{\sigma}^2)$. Supplementary Section S1 provides a complete derivation of the procedure; here we summarize it briefly.

We start by defining a family of candidate approximate posterior distributions, each of which has the same partially factorized structure:

$$q(\theta,x) = \prod_{j=1}^{M}q(x_j)\prod_{i=1}^{N}q(\theta_i). \tag{8}$$

In Equation (8), each $q(x_j)$ approximates the posterior distribution of the jth gene's values across all subtype signatures. Each $q(\theta_i)$ approximates the posterior distribution of the ith sample's subtype-mixing weights. This family of approximate posteriors imposes mutual independence between the x_j's and θ_i's. We do not use a fully factorized 'mean-field' approximation, which would treat the components within each x_j and θ_i as independent as well. Factorizing the variational distribution $q(x)$ across features allows the algorithm to update each K-dimensional feature distribution $q(x_j)$ in parallel while allowing $q(x_j)$ to incorporate dependencies across subtypes. Factorizing the variational distribution across samples in $q(\theta_i)$ accomplishes the same goals.

Next, we use a variational expectation-maximization (EM) procedure (Jordan *et al.*, 1999) to choose the best approximate posterior in the family just described. The procedure also produces approximate maximum-likelihood estimates of the hyperparameters $\hat{\phi} := \{\hat{\alpha},\hat{\lambda},\hat{\sigma}^2\}$. Variational EM maximizes a global lower bound $\mathcal{L}(q,\phi)$ on the true likelihood, by alternating between maximization over q (the variational E-step) and maximization over ϕ (the M-step).

The variational E-step is carried out via coordinate ascent: we cycle through the factors $q(x_j)$ and $q(\theta_i)$, maximizing in turn over each one with all others held fixed, until no more progress on $\mathcal{L}(q,\phi)$ can be made. The optimal coordinate update for $q(x_j)$ is known, but intractable to compute (Bishop, 2006). So, following Wang and Blei (2013), we invoke the idea of the Laplace approximation to find a good $q(x_j)$ in the family of normal distributions. The Laplace approximation is not related to the Laplace prior. For $q(\theta)$, whose optimal update is also intractable, we choose as an approximation a Dirichlet distribution, as in latent Dirichlet allocation (Blei *et al.*, 2003).

The M-step involves closed-form updates for $\hat{\lambda}$ and $\hat{\sigma}^2$, with numerical optimization applied for $\hat{\alpha}$. Algorithm 1 summarizes the GLAD inference procedure in pseudocode.

Algorithm 1 GLAD variational Laplace inference

```
Initialize q(x, θ) and φ̂
repeat
  repeat
    for j = 1 to M do
      //f(x_j, φ̂) is E_{-j}[log p(y, x_j, x_{-j}, θ)].
        (See Supplementary Section S1)
      Set x̂_j ← arg max_{x_j} f(x_j, φ̂)
      Approximate q(x_j) ≈ N(x̂_j, −∇²f(x̂_j, φ̂))
    end for
    for i = 1 to N do
      Optimize L(q, φ̂) over q(θ_i; γ_i) = Dir(γ_i)
    end for
  until change in L(q, φ̂) is small
  Set φ̂ ← arg max_φ L(q, φ)
until change in L(q, φ̂) is small
```

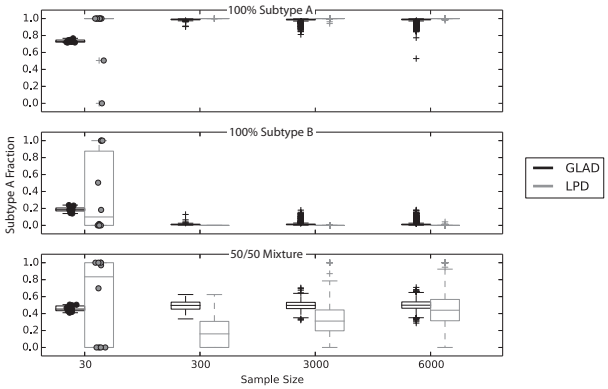

Fig. 2. Comparison of sample classification by LPD and GLAD for a range of sample sizes. Data points for $N = 30$ dataset shown for clarity

3.1 Estimating the number of subtypes (K)

Choosing the number of clusters for any clustering method is challenging (Milligan and Cooper, 1985). Methods include the silhouette measure (Rousseeuw, 1987), information theoretic measures (Sugar and James, 2003), and cross-validation (Wang, 2010). Non-parametric Bayes methods enable posterior inference about K (Teh *et al.*, 2006). Prediction accuracy on held-out data has been used in the context of document topic models (Erosheva *et al.*, 2004). Here, we use the Bayesian information criterion (BIC) (Hansen and Yu, 2001; Schwarz, 1978), choosing K to minimize the following:

$$\mathrm{BIC}(q,\phi) = -2\cdot\mathcal{L}(q,\phi)+(K\cdot(M+N+1)+2)\log(N). \quad (9)$$

We count one parameter for each element of x and θ, as well as for λ, σ^2 and α.

We expect that this method, although simple, should guide selection toward a reasonable value of K. As it is intractable to evaluate the exact likelihood, we instead use the maximized variational lower bound. We show results of this metric on tissue mixture data in Section 4.2.

4 RESULTS

4.1 Simulation study

We used simulated data with known parameter values to assess the performance of GLAD as sample size is increased, in comparison to LPD (Rogers *et al.*, 2005).

We constructed a suite of simulations as follows. We set the number of subtypes to two (denoted 'A' and 'B'), and the number of features to 500. We used sample sizes $N = 30$, 300, 3000 and 6000. In all simulations, the expected value for the first 20 features is $+2$ for subtype A and -2 for subtype B; the remaining features have an expected value of zero for both subtypes. For each simulated dataset, the first $N/3$ samples are 100% subtype A, the next $N/3$ samples are 100% subtype B and the last $N/3$ samples are a 50/50 mixture of the two subtypes. Thus, to simulate a dataset, we drew observations independently from a Gaussian distribution with mean parameter correspondingly set to subtype A's signature, subtype B's, or the average of the two. We set the Gaussian standard deviation to one.

We fit GLAD and LPD to each simulated dataset and estimated the subtype proportions for each sample. Figure 2 shows boxplots of the estimated fraction of subtype A in each sample. The sizes of the datasets are on the x-axis. The three panels in the figure correspond to the three different kinds of sample in each dataset: pure subtype A (top), pure subtype B (middle) and 50/50 mix (bottom). The boxplots show the distribution over samples of $E(\theta_{Ai} \mid y_i)$, the posterior mean fraction of subtype A detected in each sample. We show the data points for the $N = 30$ dataset because each boxplot is constructed from only 10 samples.

Figure 2 shows that in pure samples, as the sample size increases, GLAD's mixing proportion estimates concentrate at the true values. For the smallest sample size of 30, LPD's mixing proportions are more accurate than GLAD's on pure samples but much less accurate on mixed samples—in real tumors, subtype mixing is liable to be the rule rather than the exception.

At $N = 30$, the GLAD posterior underestimates the amount of subtype A for the 100% A samples and overestimates it for the 100% B samples as a consequence of the Dirichlet prior for small N. But GLAD improves quickly with more data; at $N = 300$, pure mixing proportions are well resolved, and this continues to hold on the larger datasets.

The bottom panel of Figure 2 shows that GLAD is accurate across a wide range of sample sizes for mixed samples. The median estimate is close to the true value of 0.5, and the spread of the estimates is small. In contrast, LPD has a large variance across a range of sample sizes, and it appears to exhibit some bias, which does not disappear even at $N = 6000$.

Both models use empirical Bayes in their estimation procedures. This causes some degree of smoothing in the predictions. The amount of smoothing owing to hyperparameter-fixing decreases as the sample size increases, as can be seen from the estimates for the pure samples in Figure 2. The GLAD posterior expectations are closer to 0.5 than they should be for small sample sizes, but the degree of smoothing decreases for larger sample sizes. Although empirical Bayes may appear to over-smooth the training-set predictions, this same smoothing often improves performance on test data (Efron, 2010).

4.2 *In vitro* tissue mixture data

Simulated data are useful to obtain performance metrics with known ground truth under idealized conditions. But simulated variability is not the same as variability in real gene expression data. Shen-Orr *et al.* (2010) performed an *in vitro* mixture experiment that yields a good dataset on which to test our algorithm, while still providing a true-positive control. They independently isolated RNA from rat lung, liver and brain tissue. Then they mixed those samples at varying fractions and measured the levels on microarrays with replicates. This dataset gives us known mixture fractions and a known number of subtypes with real gene expression microarray data variability.

The initial dataset contains 42 samples and 31 099 features. The measurements are the RMA-normalized microarray intensities on a $\log_2$ scale. We preprocessed the probes to select only those that have a coefficient of variation >20%, yielding a data matrix that is 1198×42 in size.

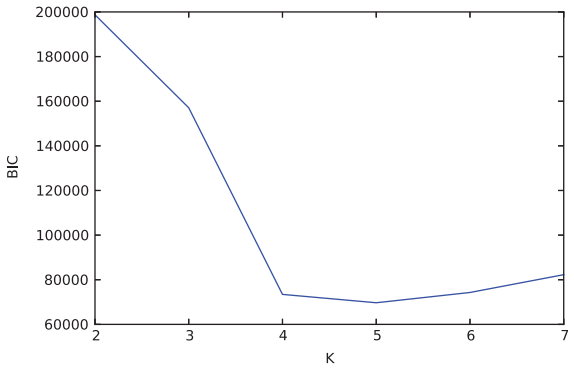

Fig. 3. BIC metric for estimating the number of clusters. The minimal BIC indicates an optimal number of clusters

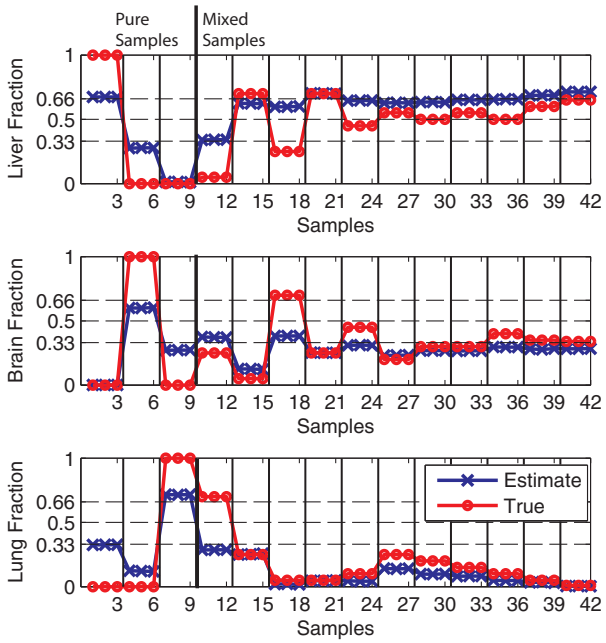

Fig. 4. Posterior estimates for mixture fractions for *in vitro* mixed tissue samples. True mixture fractions are shown as circle markers, and estimated mixture fractions are shown as cross markers, for each tissue type across 42 samples (x-axis)

We used the BIC approach to estimate K. Figure 3 shows that the BIC method chooses a value of $K = 5$, roughly in keeping with the correct value of $K = 3$.

For the remainder, we set $K = 3$, the true value, and ran variational inference for GLAD to focus the results on the inference algorithm and model. An analysis of the results for $K = 5$ is provided in Supplementary Section S2. The results are qualitatively similar for both $K = 3$ and $K = 5$. Figure 4 shows true values of θ and point estimates $\hat{\theta}$ for the three subtypes, across all 42 samples. GLAD evidently identifies both pure and mixed samples in the dataset.

This dataset has many samples containing higher fractions of brain tissue than the other tissue types. The empirical Bayes estimate of α reflects this, which in turn influences the subtype

Liver	Brain	Lung
Itih4	Mbp	LOC683399
Fga	Vsnl1	Cyp2b1 /// Cyp2b2
A1i3 /// Mug1	Snap25	Slc34a2
Gc	Plp1	Retnla
A1bg	Ptgds	Cyp4b1
Fabp1	Mbp	S100a6
Fgb	Gabra1	Sftpd
Hpd	S100b	Mfap4
Pzp	Aqp4	Napsa
Ambp	Slc6a17	Sftpa1

Fig. 5. Top 10 genes in biomarker signature for each subtype (tissue). The top genes correspond to the function of the cells in each tissue subtype

fraction estimates for the pure samples. In applications where one would prefer uniform subtype predictions a priori, α can be fixed to a scalar multiple of the ones vector. Fixing α to small positive values reduces the regularization, and thus, the underestimation of pure samples (Supplementary Section S5). However, the estimated α values do provide information about the distribution of subtypes across the entire dataset.

Many of the non-zero genes within each subtype signature are known to be associated with the corresponding tissue type. Figure 5 shows the top 10 genes ranked by the absolute value of x_j for the three subtypes. All of the top 10 genes are over-expressed in the subtype indicated. We have labeled each subtype with the most closely matching tissue type. In the liver signature, there are genes associated with blood components (ITIH4, FGA, MUG1, FGB and PZP) and genes that are associated with liver regeneration (A1BG) and molecule uptake (FABP1). In the brain subtype, the top component is MBP—a myelin sheath component. The signature also includes VSNL1—a neuronal calcium sensor, SNAP25—a synaptic vesicle dicing protein and PLP1—a transmembrane myelin protein. Finally, the lung signature contains SLC34A2, which if defective causes pulmonary alveolar microlithiasis and SFTPD, which plays a role in defense response in the lung.

4.3 TCGA GBM tumor classification

We applied the GLAD model to microarray measurements of RNA expression levels in GBM tumors obtained as part of The Cancer Genome Atlas (TCGA) project (McLendon *et al.*, 2008; Verhaak *et al.*, 2010). We used the unified filtered sample matrix provided in the Supplementary Materials (Verhaak *et al.*, 2010). These data have been normalized and filtered for the most differentially expressed genes leaving a data matrix with 1740 features and 202 primary tumor samples.

Previous work showed four canonical subtypes: classical (CL), proneural (PN), neural (NL) and mesenchymal (MES) (Verhaak *et al.*, 2010). As these are biopsy samples, we expect that there is some normal cell contamination in the samples. Normal contamination was also observed by Verhaak *et al.* (2010) and clustered within the NL subtypes. As such, we expect that mixture fractions of the NL subtype may represent normal cell contamination in our model.

We applied the BIC model selection method to this dataset and found the optimal value of $K = 6$. Briefly, one subtype is enriched for gene in the immune response process and another

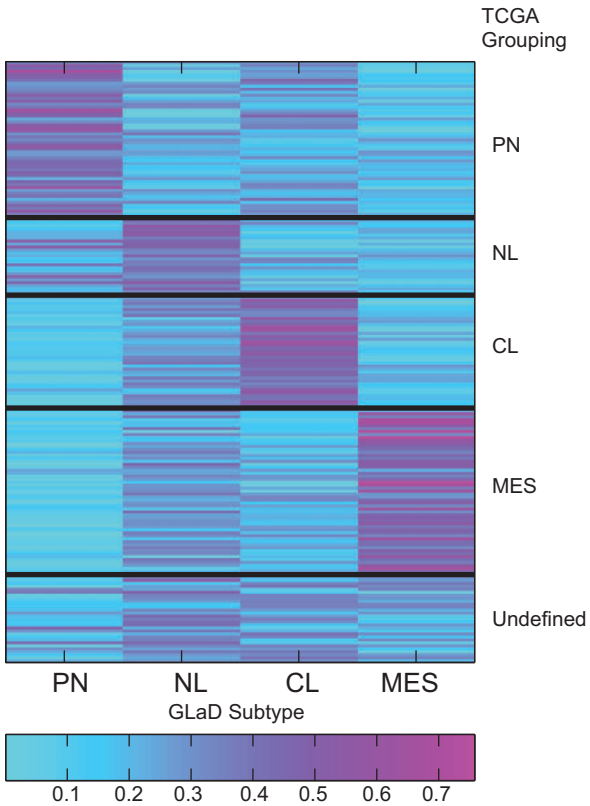

Fig. 6. Subtype distributions for each of 202 GBM samples. The estimated subtype (columns) for each sample (rows) corresponds with the assessment of the TCGA group. The GLAD algorithm also classified samples that the previous analysis left undefined

Mesynchemal (MES)	Classical (CL)	Proneural (PN)	Neural-like (NL)
DKK1	EVI2A	POSTN	**AGXT2L1**
AGXT2L1	MOG	**TMSL8**	SLCO1C1
POSTN	VSNL1	CHI3L1	AQP4
ATP1A2	IGFBP2	**DCX**	LGI1
F13A1	PLP1	LTF	ATP1A2
SCG3	ENPP2	**GPR17**	BBOX1
CD163	**EGFR**	PDPN	GABRB1
EGFR	MOBP	**PCDH11Y**	ASPA
LOX	AGXT2L1	EGFR	LMO3
COL1A1	TF	MEOX2	ALDH1L1
CKB	MBP	MOXD1	TMSL8
NCAN	DYNC1I1	PTX3	TOP2A
C1orf61	MS4A4A	EMP3	CENPF
COL1A2	UGT8	C21orf62	PLP1
MMP7	CDR1	**UGT8** (NL)	CX3CR1

Fig. 7. Sparse subtype signatures for GBM data. Upregulated genes are shown in a lighter shade (green), and downregulated genes are in a darker shade (blue). Genes identified by the TCGA analysis are shown in bold. There is high concordance between this sparse subtype signature and the more dense TCGA signatures (color online)

shows genes enriched for cell cycle processes. These findings correspond to analysis by Yoshihara *et al.* (2013) using a supervised learning approach to estimate tumor purity. A full analysis with $K = 6$ is reported in the Supplementary Sections S3 and S4. Here, we proceed with $K = 4$ to directly compare our results with those obtained by Verhaak *et al.* (2010).

Figure 6 shows the distribution over subtypes for each sample in the GBM dataset. We have labeled the subtypes based on their similarity to the reported subtypes in Verhaak *et al.* (2010). They found that the two normal samples in their dataset were classified as NL. As expected, there is a considerable mixing between the NL subtype and other subtypes in our analysis, possibly due to normal cell contamination. We also see significant mixing among the other subtypes, indicating that the samples are indeed generally not pure.

Figure 7 shows the genes involved in the four subtypes identified by GLAD. Like topic models, each subtype has coefficient values for all features. However, the Laplace prior has the effect of regularizing uninformative coefficients toward zero, yielding a sparse signature. We report only the top 15 candidates for the signature here; the full table is available in the supplementary information. Genes in green are upregulated in the subtype, and genes in blue are downregulated. Genes in bold were also identified by Verhaak *et al.* (2010) for that subtype, except UGT8, which was associated with NL-like, whereas we associated it with PN.

Upregulation of DKK1 and POSTN is associated with the MES subtype in our analysis. DKK1 plays a role in inhibition of the Wnt signaling pathway and the presence of bone lesions in multiple myeloma patients (Tian *et al.*, 2003). Although the entire function of perstonin (POSTN) is not known, it is frequently upregulated in cancers and has been associated with adhesion and differentiation of osteoblasts (Kudo *et al.*, 2007). These markers, in combination with collagen-specific genes, COL1A1 and COL1A2, point to the role of these genes in the precursor to bone and cartilage development for this subtype's MES-like genomic character.

The classical subtype has only one highly upregulated gene—EGFR. Verhaak *et al.* (2010) observed this association at both the mRNA and DNA levels in 97% of classical-like cells indicating its strong association. The PN actin and microtubule-related genes TMSL8 and DCX are over-expressed in the PN subtype (Brown *et al.*, 2003). Finally, the NL-like subtype contains the central nervous system water channel AQP4 and other genes associated with normal NL cells.

5 DISCUSSION

We have presented GLAD, a new statistical model for the genomic classification of solid tumors. The model does not make an all-or-none classification for each sample; instead, it estimates a per-sample distribution over subtypes. Simultaneously, the model estimates a sparse expression signature for each subtype. We provide an interpretation of the model as a regularized form of matrix factorization.

We compared GLAD with LPD, another clustering and classification model, using simulated data. At small dataset sizes, we found that GLAD overestimates the amount of mixing for pure samples, whereas LPD is quite imprecise for any particular sample; in our view, one should expect heterogeneity to be the norm in applications. As we increased the simulated sample size, the performance of GLAD improved quickly, owing to the decreased weight of the prior in the inference about θ_i. GLAD is able to identify gene expression patterns associated with different tissue types from real microarray data, even when the sample is a

heterogeneous mixture. An analysis of TCGA GBM data shows that there is significant heterogeneity in those clinical samples.

Our Laplace variational inference algorithm provides a fast method for statistical inference using GLAD. However, it is an approximation. A Markov-Chain Monte Carlo approach has potential to reduce bias, at the expense of possibly much higher computational cost. Our variational approximation is also one of many possible factorizations of the model. We are actively testing other ways of approximating the posterior distribution that retain the computational advantages while improving the quality of the approximation.

When fitting GLAD, we optimize a non-convex objective function by coordinate ascent. As such, we are only able to identify a local maximum in the variational auxiliary function, not necessarily a global maximum. Restarting the algorithm using random initializations can help, and we found it important to do so.

Estimating the number of clusters, K, is a universal problem for clustering algorithms. Our approach, using the BIC, provides only a rough guideline. However, this approach may require much more data for accurate parameter estimation than needed in the current model. Another alternative is to choose K by cross-validation. This shifts the problem from choosing K to choosing an appropriate cross-validatory accuracy metric.

Recent studies in breast cancer (Curtis *et al.*, 2012) and colon cancer (De Sousa E Melo *et al.*, 2013) have suggested different subtypes, which show distinct survival outcomes. As more samples are analyzed a clearer picture of cancer subtypes and number of subtypes develops.

We can extend GLAD in several directions. Other data types such as imaging, DNA copy-number variation and DNA methylation may be incorporated as additional x and y outcomes. Structured sparsity may be incorporated into the λ hyperparameter. Finally, a hierarchical Dirichlet process may be incorporated into θ to learn the number of subtypes K.

ACKNOWLEDGEMENTS

The authors would like to thank Sia Najafi for computational support. P.F. and H.S. are supported by seed funding from Worcester Polytechnic Institute.

Conflict of interest: none declared.

REFERENCES

Airoldi,E.M. *et al.* (2008) Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, **9**, 1981–2014.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., Secaucus, NJ, USA.

Blei,D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Bonavia,R. *et al.* (2011) Heterogeneity maintenance in glioblastoma: a social network. *Cancer Res.*, **71**, 4055–4060.

Brown,J.P. *et al.* (2003) Transient expression of doublecortin during adult neurogenesis. *J. Comp. Neurol.*, **467**, 1–10.

Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

De Sousa E Melo,F. *et al.* (2013) Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.*, **19**, 614–618.

Dexter,D.L. *et al.* (1978) Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res.*, **38**, 3174–3181.

Efron,B. (2010) *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University press, Cambridge, UK.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Erosheva,E. *et al.* (2004) Mixed-membership models of scientific publications. *Proc. Natl Acad. Sci. USA*, **101** (**Suppl**), 5220–5227.

Falush,D. *et al.* (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes*, **7**, 574–578.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**RESEARCH0059.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366**, 883–892.

Hansen,M.H. and Yu,B. (2001) Model selection and the principle of minimum description length. *J. Am. Stat. Assoc.*, **96**, 746–774.

Heppner,G.G.H. (1984) Tumor heterogeneity. *Cancer Res.*, **44**, 2259–2265.

Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.

Hoyer,P. (2004) Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Jordan,M.I. *et al.* (1999) An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

Kabán,A. (2007) On bayesian classification with laplace priors. *Pattern Recognit. Lett.*, **28**, 1271–1282.

Koboldt,D.C. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Kudo,Y. *et al.* (2007) Periostin: novel diagnostic and therapeutic target for cancer. *Histol. Histopathol.*, **22**, 1167–1174.

Lee,H. *et al.* (2006) Efficient sparse coding algorithms. *Adv. Neural Inf. Process. Syst.*

MacKay,D.J.C. (1992) Bayesian interpolation. *Neural. Comput.*, **4**, 415–447.

McLendon,R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Milligan,G.W. and Cooper,M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Parker,J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Parsons,D.W. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.

Rogers,S. *et al.* (2005) The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 143–156.

Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Shen-Orr,S.S. *et al.* (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Singh,A.P. and Gordon,G.J. (2008) *Machine Learning and Knowledge Discovery in Databases*. volume 5212 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sugar,C.A. and James,G.M. (2003) Finding the number of clusters in a dataset. *J. Am. Stat. Assoc.*, **98**, 750–763.

Taddy,M. (2012) Multinomial inverse regression for text analysis. *J. Am. Stat. Assoc.*, **108**, 755–770.

Teh,Y.W. *et al.* (2006) Hierarchical dirichlet processes. *J. Am. Stat. Assoc.*, **101**, 1566–1581.

Tian,E. *et al.* (2003) The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *N. Engl. J. Med.*, **349**, 2483–2494.

Verhaak,R.G.W. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.

Wang,C. and Blei,D.M. (2009) Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *Adv. Neural Inf. Process. Syst.*, 1982–1989.
Wang,C. and Blei,D.M. (2013) Variational inference in nonconjugate models. *J. Mach. Learn. Res.*, **14**, 1005–1031.
Wang,J. (2010) Consistent selection of the number of clusters via crossvalidation. *Biometrika*, **97**, 893–904.
Wang,X. and McCallum,A. (2006) *Topics Over Time: a Non-Markov Continuous-time Model of Topical Trends. A Non-Markov Continuous-time Model of Topical trends*. ACM, New York, NY, USA.
Yoshihara,K. *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.