Algorithms for Regularized Linear Discriminant Analysis

Jan Kalina[1] and Jurjen Duintjer Tebbens[2,3]

[1]*Dept. of Medical Informatics and Biostatistics, Institute of Computer Science AS CR, Prague, Czech Republic*

[2]*Dept. of Computational Methods, Institute of Computer Science AS CR, Prague, Czech Republic*

[3]*Dept. of Biophysics and Physical Chemistry, Faculty of Pharmacy in Hradec Králové, Charles University in Prague, Hradec Králové, Czech Republic*

kalina@cs.cas.cz, duintjertebbens@cs.cas.cz

Keywords: Classification analysis, Regularization, High-dimensional data, Matrix decomposition, Computational aspects.

Abstract: This paper is focused on regularized versions of classification analysis and their computation for high-dimensional data. A variety of regularized classification methods has been proposed and we critically discuss their computational aspects. We formulate several new algorithms for regularized linear discriminant analysis, which exploits a regularized covariance matrix estimator towards a regular target matrix. Numerical linear algebra considerations are used to propose tailor-made algorithms for specific choices of the target matrix. Further, we arrive at proposing a new classification method based on L_2-regularization of group means and the pooled covariance matrix and accompany it by an efficient algorithm for its computation.

1 INTRODUCTION

Classification analysis methods have the aim to construct (learn) a decision rule based on a training data set, which is able to automatically assign new data to one of K groups. Linear discriminant analysis (LDA) is a standard statistical classification method. In the whole paper, we consider n observations with p variables, observed in K different samples (groups) with $p > K \geq 2$,

$$X_{11}, \ldots, X_{1n_1}, \ldots, X_{K1}, \ldots, X_{Kn_K}, \qquad (1)$$

where $n = \sum_{k=1}^{K} n_k$. LDA assumes the data in each group to come from a Gaussian distribution. The covariance matrix Σ is assumed to be the same across groups and its estimator will be denoted by S.

Various tasks in bioinformatics deal with high-dimensional data, i.e. data with the number of observed variables p exceeding the number of observations. Analyzing data in this so-called large p/small n problem is especially important e.g. in gene expression studies. Unfortunately, LDA in its standard form is infeasible for $n < p$, because the matrix S of size p is singular and computing its inverse must be replaced by an appropriate alternative. Available approaches in this context are based e.g. on pseudoinverse matrices, which are however unstable due to a small n (Guo et al., 2007).

Regularized versions of LDA for $n \ll p$ were proposed (Guo et al., 2007) mainly for applications in bioinformatics. They are based on such regularized estimator of the covariance matrix, which is guaranteed to be regular and positive definite even for $n \ll p$. Fast computation and numerical stability is a key requirement expected from reliable statistical and data mining procedures (Kogan, 2007; Duintjer Tebbens and Schlesinger, 2007) and remains to be an important issue also for regularized classification methods (Hastie et al., 2008; Pourahmadi, 2013). Let us now describe the most important approaches and critically discuss their possible computation.

This paper studies efficient algorithms for computing various regularized versions of LDA. Section 2 of this paper formulates several algorithms for regularized LDA, which exploits a regularized covariance matrix estimator towards a regular target matrix. The computational effectivity of the algorithms is inspected using arguments of numerical linear algebra. For a specific choice of the target matrix, we are able to propose a tailor-made algorithm with a lower computational cost compared to algorithms which are formulated for a general context. Besides, we arrive at proposing a new version of LDA based on regularization in the sense of the L_2 norm and accompany it by an efficient algorithm for its computation in Section 3. The classification performance of the methods is illustrated on real data in Section 4. Finally, Section 5 concludes the paper.

2 COMPUTATION OF REGULARIZED LINEAR DISCRIMINANT ANALYSIS

This section is devoted to proposing and comparing new algorithms for a habitually used version of the regularized LDA. We use suitable matrix decompositions to propose efficient algorithms either for a general choice of T or for its specific choices. To the best of our knowledge, tailor-made algorithms for a specific T have not been described. We compare the new algorithms in terms of their computational costs as well as numerical stability.

We will describe one of habitually used regularized versions of LDA. This will be denoted as LDA* to avoid confusion, because the concept of regularized discriminant analysis encompasses several different methods (Pourahmadi, 2013; Kalina, 2014) A given target matrix T will be used, which must be a regular symmetric positive definite matrix of size $p \times p$. Its most common choices include the identity matrix I_p or a diagonal (non-identity) matrix; other target matrices have been considered by (Schäfer and Strimmer, 2005).

Let us denote the mean of the observed values in the k-th group ($k = 1, \ldots, K$) by $\bar{X}_k$. LDA* assigns a new observation $Z = (Z_1, \ldots, Z_p)^T$ to group k, if $l_k^* > l_j^*$ for every $j \neq k$, where the regularized linear discriminant score for the k-th group ($k = 1, \ldots, K$) has the form

$$l_k^* = \bar{X}_k^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k^T (S^*)^{-1} \bar{X}_k + \log p_k, \qquad (2)$$

where p_k is a prior probability of observing an observation from the k-th group and

$$S^* = \lambda S + (1-\lambda) T \qquad (3)$$

for $\lambda \in (0,1]$ denotes a regularized estimator of the covariance matrix across groups. We do not treat the situation with $l_k^* = l_{k'}^*$ for $k' \neq k$ separately, because it occurs with a zero probability for data coming from a continuous distribution. Equivalently, LDA* assigns a new observation Z to group k, if

$$(\bar{X}_k - Z)^T S^{*-1} (\bar{X}_k - Z) = \\ = \min_{j=1,\ldots,K} \left\{ (\bar{X}_j - Z)^T S^{*-1} (\bar{X}_j - Z) \right\}. \qquad (4)$$

Various versions of the target matrix T have been suggested by (Schäfer and Strimmer, 2005). We point out that the formula (3) can be justified by Bayesian reasoning. Let us assume normally distributed data with a covariance matrix Σ, while Σ^{-1} is assumed to be a random variable following Wishart distribution $W_p(T/\gamma, k)$ for some $\gamma > 0$ and an integer k. Following (Haff, 1980), the mean of the posterior distribution of Σ is equal to (3) up to a normalizing constant, which does not influence the classification rule.

First, the standard approach for computing LDA* may be improved by employing the eigendecomposition of S^* for a fixed λ. A suitable value of λ is found by a cross-validation in the form of a grid search over all possible values of $\lambda \in (0,1]$.

Algorithm 1. *LDA* for the general regularization (3) based on eigendecomposition.*

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \ldots, \bar{X}_K - Z] \qquad (5)$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$.

Step 2 *Compute S^* according to (3) with a fixed $\lambda \in (0,1]$.*

Step 3 *Compute the eigendecomposition of S^* as*

$$S^* = Q_* D_* Q_*^T. \qquad (6)$$

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q_*^T A \qquad (7)$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.*

The group assignment (4) is done by using

$$(\bar{X}_j - Z)^T S^{*-1} (\bar{X}_j - Z) = \\ = (\bar{X}_j - Z)^T Q_* D_*^{-1} Q_*^T (\bar{X}_j - Z) = \\ = \|D_*^{-1/2} Q_*^T (\bar{X}_j - Z)\|^2. \qquad (8)$$

The costs of the algorithm can be made reduced by replacing the eigendecomposition of S^* with its Cholesky decomposition

$$S^* = L_* L_*^T, \qquad (9)$$

where L_* is a nonsingular lower triangular matrix.

Algorithm 2. *LDA* for the general regularization (3) based on Cholesky decomposition.*

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \ldots, \bar{X}_K - Z] \qquad (10)$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$.

Step 2 *Compute S^* according to (3) with a fixed $\lambda \in (0,1]$.*

Step 3 *Compute the Cholesky factor L_* of S^*.*

Step 4 *Compute the matrix*

$$B = L_*^{-T} A \tag{11}$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.*

For specific target matrices, we can further reduce computational costs by using the following algorithm for LDA*. The pooled estimator S can be written in the form $S = Y^T Y$, where

$$\begin{aligned} Y = [&X_{11} - \bar{X}, \ldots, X_{1n_1} - \bar{X}, \ldots, \\ &X_{K1} - \bar{X}, \ldots, X_{Kn_K} - \bar{X}]^T \end{aligned} \tag{12}$$

is of size $n \times p$. Then using the singular value decomposition (SVD) of Y in the form

$$Y = P \Sigma Q^T, \tag{13}$$

we can express the eigendecomposition of S as

$$S = Y^T Y = (P \Sigma Q^T)^T P \Sigma Q^T = Q \Sigma^2 Q^T. \tag{14}$$

The costs will be about $4 \cdot np^2$ floating point operations, thus with $p \gg n$ the gain is considerable.

Moreover, if

$$S^* = \lambda S + (1-\lambda) I_p, \quad \lambda \in (0,1], \tag{15}$$

we immediately obtain the needed eigendecomposition of S^* as

$$S^* = \lambda S + (1-\lambda) I_p = Q \left(\lambda \Sigma^2 + (1-\lambda) I_p \right) Q^T. \tag{16}$$

The SVD can be computed in a backward stable way with all singular values accurate up to machine precision level (Barlow et al., 2005). For the special case (15), which is commonly denoted as Tikhonov or ridge regularization of S, a more efficient computation can be performed as follows.

Algorithm 3. *LDA* for the ridge regularization (15).*

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \ldots, \bar{X}_K - Z] \tag{17}$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$ and compute the matrix Y in (12).

Step 2 *Compute the singular value decomposition of Y as*

$$Y = P \Sigma Q^T, \tag{18}$$

with singular values $\{\sigma_1, \ldots, \sigma_n\}$ and complement these singular values with $p-n$ zero values $\sigma_{n+1} = \cdots = \sigma_p = 0$.

Step 3 *For a fixed $\lambda \in (0,1]$, compute $D_* =$*

$$diag\{\lambda \sigma_1^2 + (1-\lambda), \ldots, \lambda \sigma_p^2 + (1-\lambda)\}. \tag{19}$$

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q^T A \tag{20}$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.*

Eigenvalues of S^* evaluated in (19) can be interpreted as regularized eigenvalues. They are however different from shrinkage eigenvalues (Pourahmadi, 2013), which are obtained by regularizing S by applying a penalization criterion directly on eigenvalues.

3 L_2-REGULARIZED LINEAR DISCRIMINANT ANALYSIS

Disadvantages of SCRDA include a computational intensity as well as an inconsistent approach to regularization. The means are namely modified by an L_1-norm regularization and the covariance matrix in the sense of the L_2-norm. Examples of other regularized LDA versions include the Prediction Analysis of Microarrays (PAM) (Tibshirani et al., 2003), which can be described as a diagonalized LDA (DLDA) with means regularized in the L_1-norm.

As an alternative, this section proposes a new regularized version of LDA denoted as L_2-LDA. As a unique feature, the means in each group as well as the pooled covariance matrix are regularized in the same way, i.e. in the L_2-norm. We propose an efficient algorithm for the computation of the method.

The classification rule of L_2-LDA assigns a new observation Z to the k-th group, if $l_k^\dagger > l_j^\dagger$ for every $j \neq k$, where

$$l_k^\dagger = \bar{X}_k'^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k'^T (S^*)^{-1} \bar{X}_k' + \log p_k \tag{21}$$

and $\bar{X}_k'$ denotes the shrunken mean of the k-th group towards the overall mean computed across groups. The method can be interpreted as based on a L_2 regularized Mahalanobis distance. As another contrast with the habitually used algorithm of SCRDA, we will estimate the parameter λ in a straightforward way using an asymptotically optimal value minimizing the mean square error (Schäfer and Strimmer, 2005). To avoid confusion, the asymptotically optimal value of λ will be denoted by $\lambda^\dagger$ and the corresponding regularized covariance matrix by

$$S^\dagger = \lambda^\dagger S + (1-\lambda^\dagger) T. \tag{22}$$

Algorithm 4. *L_2-LDA.*

Step 1 *Compute $\lambda^\dagger$ as*

$$\lambda^\dagger = \frac{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}\widehat{var}(S_{ij})}{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}S_{ij}^2+\sum_{i=1}^{p}(S_{ii}-1)^2}, \qquad (23)$$

where $\widehat{var}(S_{ij})$ is the maximum likelihood estimator of the variance of values S_{ij} for a fixed i and j.

Step 2 *Compute the eigendecomposition of $S^\dagger$ as*

$$S^\dagger = Q_* D_* Q_*^T. \qquad (24)$$

Step 3 *For a fixed $\delta \in [0,1]$, compute*

$$\bar{X}_k' = \delta\bar{X}_k + (1-\delta)\bar{X}, \qquad k = 1,\ldots,K. \qquad (25)$$

Step 4 *Assign Z to group k, if*

$$\|D_*^{-1/2}Q_*^T(\bar{X}_k'-Z)\| = \\ = \min_{j=1,\ldots,K}\|D_*^{-1/2}Q_*^T(\bar{X}_j'-Z)\|. \qquad (26)$$

Step 5 *Repeat steps 3 and 4 for various δ and find the optimal classification rule yielding the best classification performance.*

The main computational costs are in step 2; the eigendecomposition costs about $9\cdot p^3$ floating point operations. Note that we need not (and should never) compute the inverse of S^*, thus avoiding additional computations of the Mahalanobis distance, which is expensive of order p^3 and numerically rather unstable.

Algorithm 4 is formulated for a general target matrix T. For a specific choice of T, a computationally cheaper method can be obtained in an analogous way as Algorithms 2 and 3 from the general Algorithm 1. Particularly, the costs of Cholesky decomposition (9) are about $p^3/3$ floating point operations. On the other hand, Cholesky decomposition will suffer from instability when S^* is not positive definite.

Analogous reasoning can be applied to obtain an algorithm for L_2-regularized version of quadratic discriminant analysis (QDA), which is another standard classification method derived for data following a Gaussian distribution, but without the assumption of a common covariance matrix. A regularized covariance matrix estimator S_k^* in the k-th group will be considered with the optimal value of the regularization parameter, again yielding the best classification performance. The regularized quadratic discriminant score for the k-th group has the form

$$q_k^* = \bar{X}_k^T(S_k^*)^{-1}Z - \frac{1}{2}\bar{X}_k^T(S_k^*)^{-1}\bar{X}_k - \\ -\frac{1}{2}Z^T(S_k^*)^{-1}Z + \frac{1}{2}\log|S_k^*| + \log p_k, \qquad (27)$$

where $|\cdot|$ denotes the determinant of a matrix.

4 EXAMPLES

We present two examples on real molecular genetic data sets in order to illustrate the behavior of the newly proposed L_2-LDA method and to compare its with performance of classical classification procedures.

Example 1 contains data from our own cardiovascular genetic study on 24 patients having a cerebrovascular stroke and 24 control persons (Kalina and Zvárová, 2013). The were $p = 38\,590$ gene transcripts measured, which correspond to the whole genome.

In Example 2, a prostate cancer metabolomic data set (Sreekumar et al., 2009) is analyzed, which contains $p = 518$ metabolites measured over two groups of patients, namely those with a benign prostate cancer (16 patients) and with other cancer types (26 patients). The task in both examples is to learn a classification rule allowing to discriminate between the two classes of individuals.

In both examples, we computed the classification methods described in this paper using the algorithms of Sections 2 and 3. For comparison, we computed also other available classification methods, including the support vector machines (SVM), a classification tree, Kohonen's self-organizing map, or a multilayer perceptron with 2 hidden layers. Various regularized versions of LDA include the most common choice $T = I_p$ or another choice

$$S^* = \lambda S + (1-\lambda)sI_p, \quad s = \sum_{i=1}^{p} S_{ii}/p, \qquad (28)$$

for $\lambda \in (0,1]$. We used the default settings to compute them in user-submitted packages, which accompany the R free software and are listed also in Table 1. The classification performance is measured by means of the Youden's index, which is defined as

$$\text{sensitivity} + \text{specificity} - 1. \qquad (29)$$

The results performed on raw data as well as after a dimensionality reduction reveal that the regularized versions of LDA perform quite similarly. The newly proposed method L_2-LDA with an efficient algorithm seems to perform comparably with the available regularized methods with less efficient computation. Besides, the choice of the target matrix T does not seem to play an important role. Some of standard classification methods are infeasible because of the dimensionality ($n \ll p$).

Further, we investigated the effect of dimensionality reduction on the classification performance. Principal component analysis (PCA) is performed and the consequent classification is applied on 20 principal components. To explain the notation, for example the

Method	S^*	*R* Package	Function	Youden's index Example 1	Example 2
SVM	-	*e1071*	*svm*	1.00	1.00
Classification tree	-	*tree*	*tree*	0.94	0.97
Self-organizing map	-	*kohonen*	*som*	0.88	0.93
Multilayer percpetron	-	*nnet*	*nnet*	Infeasible	Infeasible
LDA	-	*MASS*	*lda*	Infeasible	Infeasible
SCRDA	(15)	*rda*	*rda*	1.00	1.00
LDA*	(15)	-	-	1.00	1.00
LDA*	(28)	-	-	1.00	1.00
L_2-LDA	(15)	-	-	1.00	1.00
L_2-LDA	(28)	-	-	1.00	1.00
PCA $\Longrightarrow$ LDA	-	-	-	0.54	0.90
PCA $\Longrightarrow$ SCRDA	(15)	-	-	0.71	0.92
PCA $\Longrightarrow$ LDA*	(15)	-	-	0.63	0.81
PCA $\Longrightarrow$ LDA*	(28)	-	-	0.63	0.81
PCA $\Longrightarrow$ L_2-LDA	(15)	-	-	0.71	0.92
PCA $\Longrightarrow$ L_2-LDA	(28)	-	-	0.71	0.92

Table 1: Results of Example 1 and Example 2. LDA* was computed using Algorithm 3 for the choice (15) and Algorithm 2 for (28). L_2-LDA was computed using Algorithm 4. PCA uses 20 principal components.

approach denoted as PCA $\Longrightarrow$ L_2-LDA corresponds to performing L_2-LDA (using Algorithm 4) on 20 principal components of the original data.

The method PCA $\Longrightarrow$ L_2-LDA yields improved results compared to its standard counterpart (PCA $\Longrightarrow$ LDA) and is not outperformed by any other method. It is remarkable that the combination PCA $\Longrightarrow$ LDA* does not improve the results compared to PCA $\Longrightarrow$ LDA. The first three principal components seem rather arbitrary and they explain only 6 %, 6 %, and 5 % of the variability in the data, respectively. Besides, we did not find any clear interpretation of the principal components and there seems no remarkable small group of genes responsible for a large portion of variability of the data.

5 CONCLUSIONS

For high-dimensional continuous data, the main obstacle of the traditional Mahalanobis distance is singularity of the empirical covariance matrix. As a solution, various regularized versions of the LDA have been proposed, which are commonly used to learn a classification rule from high-dimensional data. This paper presents our view that the methodology in its standard form represents a set of ad hoc procedures rather than a coherent approach, moreover without efficient algorithms available for the computation. Besides, we explain some open problems concerning regularized versions of LDA, e.g. specific algorithms tailor-made for a specific choice of the target matrix.

Several new algorithms are proposed for a regularized LDA in Section 2. We propose the L_2-regularized version of LDA, advocating our position that the mean and covariance matrix of multivariate data can be estimated by means of the same regularization principle. A new regularized classification method L_2-LDA is proposed in Section 3 and accompanied by an efficient algorithm. Its computational costs are discussed. The regularization can be interpreted as a tailor-made correction for a small sample size. If $n < p$ for larger sample sizes, the effect of the regularization will become of a smaller importance.

L_2-LDA can be interpreted as a shrinkage estimator of the covariance matrix, in the light of the Stein's result of estimating the mean of multivariate normal data (Stein, 1956; Hastie et al., 2008). It is possible to interpret the method as an approach based on a regularized version of the Mahalanobis distance. The method is reliable under an implicit assumption that the variability is not substantially different across variables. The optimal regularization parameters seem to be reasonable in the classification analysis context. Another possibility is to regularize the within-group covariance matrix instead of regularizing S, which is however computationally more intensive. An analysis of two real data sets reveals its classification performance to be comparable to available regularized classification methods for high-dimensional data.

In general, some regularized statistical methods for the analysis of high-dimensional data have been empirically observed to possess reasonable robust-

ness properties with respect to outlying measurement in the data. To give an example, regularized means has been observed to cause a certain local robustness against small departures in the observed data (Tibshirani et al., 2003). Regularization itself cannot ensure robustness against serious outliers (Filzmoser and Todorov, 2011) for continuous data. In the words of robust statistics, regularization does not imply a robustness in terms of the breakdown point and regularized LDA cannot replace robust classification procedures with a high breakdown point (Kalina, 2012). It remains an open problem to investigate systematically the relationship between regularization and statistical robustness for continuous data. Another warning should be given that there is no reason to suppose that the optimal procedure for the regularized model will perform well away from that model (Davies, 2014).

Alternative approaches could be formulated by means of regularization requiring a certain level of sparsity (Chen et al., 2012). Moreover, L_2-LDA can be derived in an alternative way as a Bayesian estimator or as the optimal method by means of robust optimization (Xanthopoulos et al., 2013).

As a future research, we plan to investigate suitable choices of the target matrix $\mathbf{T}$ and extend the regularized Mahalanobis distance to the context of cluster analysis. From the theoretical point of view, robustness of LDA regularized in the L_1-norm has not been inspected as well as regularized versions of the highly robust MWCD estimator (Kalina, 2012). We plan to apply and compare regularized versions of LDA to pattern recognition problems in the analysis of 3D neuroimages of spontaneous brain activity. There, we plan to exploit the new L_2-LDA without the usual sparseness assumption, allowing to choose $\mathbf{T}$ to model the high correlation of neighboring voxels.

ACKNOWLEDGEMENTS

The work was financially supported by the Neuron Fund for Support of Science. The work of J. Kalina was supported by the grant GA13-17187S of the Czech Science Foundation. The work of J. Duintjer Tebbens was supported by the grant GA13-06684S of the Czech Science Foundation.

REFERENCES

Barlow, J., Bosner, N., and Drmac, Z. (2005). A new stable bidiagonal reduction algorithm. *Linear Algebra and its Applications*, 397:35–84.

Chen, X., Kim, Y., and Wang, Z. (2012). Efficient minimax estimation of a class of high-dimensional sparse precision matrices. *IEEE Transactions on Signal Processing*, 60:2899–2912.

Davies, P. (2014). *Data Analysis and Approximate Models: Model Choice, Location-Scale, Analysis of Variance, Nonparametric Regression and Image Analysis*. Chapman & Hall/CRC, Boca Raton.

Duintjer Tebbens, J. and Schlesinger, P. (2007). Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Computational Statistics & Data Analysis*, 52:423–437.

Filzmoser, P. and Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chinica Acta*, 705:2–14.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.

Haff, L. (1980). Empirical bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 1980:586–597.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The elements of statistical learning*. Springer, New York, 2nd edition.

Kalina, J. (2012). Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32(2):3–16.

Kalina, J. (2014). Classification analysis methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34:10–18.

Kalina, J. and Zvárová, J. (2013). Decision support systems in the process of improving patient safety. In *E-health Technologies and Improving Patient Safety: Exploring Organizational Factors*, pages 71–83. IGI Global, Hershey.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press, Cambridge.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. Wiley, New York.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 32:1–30.

Sreekumar et al., A. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–206.

Tibshirani, R., Hastie, T., and Narasimhan, B. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 18:104–117.

Xanthopoulos, P., Pardalos, P., and Trafalis, T. (2013). *Robust data mining*. Springer, New York.