# Geometric optimization methods for the analysis of gene expression data

M. Journée[1], A. E. Teschendorff[2], P.-A. Absil[3], S. Tavaré[2], and R. Sepulchre[1]

[1] Department of Electrical Engineering and Computer Science, University of Liège, Belgium.
[2] Breast Cancer Functional Genomics Program, Cancer Research UK Cambridge Research Institute, Department of Oncology, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK.
[3] Department of Mathematical Engineering, Université catholique de Louvain, Belgium.

MJ and AET contributed equally to this work.

## Abstract

DNA microarrays provide such a huge amount of data that unsupervised methods are required to reduce the dimension of the data set and to extract meaningful biological information. This work shows that Independent Component Analysis (ICA) is a promising approach for the analysis of genome-wide transcriptomic data. The paper first presents an overview of the most popular algorithms to perform ICA. These algorithms are then applied on a microarray breast-cancer data set. Some issues about the application of ICA and the evaluation of biological relevance of the results are discussed. This study indicates that ICA significantly outperforms Principal Component Analysis (PCA).

## 1 Introduction

The transcriptome is the set of all mRNA molecules in a given cell. Unlike the genome, which is roughly similar for all the cells of an organism, the transcriptome may vary from one cell to another according to the biological functions of that cell as well as to the external stimuli. The transcriptome reflects the activity of all the genes within the cell. The quantity of a given mRNA is determined by a complex interaction between cooperative and counteracting biological processes. Understanding the intricate mechanism that induces the

mRNA expression is an important step in elucidating the relation between the transcriptome of a cell and its phenotype. Microarray technology provides a quantitative measure of the concentration of mRNA molecules in a cell for the whole transcriptome in a systematic way. This measure is called the expression level of a gene. We refer to [RCTH05] and references therein for more details about microarrays.

Microarray technology provides a huge amount of data, typically related to several thousand genes over a few hundred experiments, which correspond, e.g., to different patients, tissues or environmental conditions. Gene expression data sets are so large that a direct interpretation of them is usually infeasible. Unsupervised methods are required to reduce the dimension of the data set and to provide some biological insight in an automatic way. A typical approach is Principal Component Analysis (PCA) that generates a linear representation of the data in terms of components that are uncorrelated [ABB03]. This linear decorrelation can reveal interesting biological information. Nevertheless, mapping the data to independent biological processes should provide a more realistic model.

The present paper proposes the use of ICA to help the biological interpretation of a gene expression data set that is related to breast cancer [WKZ$^+$05]. The ICA model is well-suited for the analysis of data sets that enclose an independence assumption within the data generation process. Intuitively, gene expression results from several biological processes (here we call them "expression modes") that take place independently. Each of these biological processes involves various biological functions and rely on the activation or inhibition of a subset of genes. Several studies have already shown the value of ICA in the gene expression context, notably Liebermeister [Lie02], who was the first to apply ICA to gene expression data. Important results on some bacterial and human databases are also detailed in [MMSM02, LB03, SHK$^+$03]. These studies identified independent components and used the Gene Ontology (add citation from our other paper.!) framework to evaluate their biological significance. The present paper extends the results from these previous studies by evaluating ICA in the framework of biological and cancer-related pathways, which constitutes the more relevant validation framework since genes with different GO-terms may be involved in the same pathway or biological process. Specically, it tests the ICA-model for gene expression by showing that ICA significantly outperforms a non-ICA based method (PCA). Furthermore, it discusses some issues about the way to apply ICA and to evaluate the results. These issues are disregarded in most of the previous studies.

This paper starts with a review of some standard algorithms to perform ICA. The link between the ICA approach and the theory of geometric optimization is first outlined (Section 2). The three fundamental components of each ICA algorithm are then discussed, namely the contrast function (Section

3), the matrix manifold (Section 4) and the optimization process (Section 5). The last part of the paper (Section 6) illustrates the application of ICA to the analysis of gene expression data.

## 2 ICA as a geometric optimization problem

The ICA approach was originally dedicated to the blind source separation problem, which recovers independent source signals from linear mixtures of them. As in the original paper of Comon [Com94], a linear instantaneous mixture model will be considered here,

$$X = AS \tag{1}$$

where $X$, $A$ and $S$ are respectively matrices in $\mathbb{R}^{n\times N}$, $\mathbb{R}^{n\times p}$ and $\mathbb{R}^{p\times N}$ with $p$ less or equal than $n$. The rows of $S$ are assumed to stand for samples of independent random variables. Thus, ICA provides a linear representation of the data $X$ in terms of components $S$ that are statistically independent. Random variables are independent if the value of any one variable does not carry any information on the value of any other variable. By definition, $p$ statistically independent random variables $s_i$ have a joint probability distribution that equals the product of their marginal distributions, i.e.,

$$p(s_1, \ldots, s_p) = p(s_1) \ldots p(s_p).$$

Each ICA algorithm is based on the inverse of the mixing model (1), namely

$$Z = W^T X,$$

where $Z$ and $W$ are respectively matrices of $\mathbb{R}^{p\times N}$ and $\mathbb{R}^{n\times p}$. The rows of $Z$ should represent random variables that are statistically independent. Unfortunately, the number of degrees of freedom available though the matrix $W$ is usually insufficient for an exact independence. Thus, the rows of $Z$ are just expected to represent random variables that are as independent as possible, such that ICA can be treated as an optimization problem.

Given an $n \times N$ data matrix $X$, an ICA algorithm aims at computing an optimum of a *contrast function*

$$\gamma : \mathbb{R}^{n\times p} \to \mathbb{R} : W \mapsto \gamma(W)$$

that estimates the statistical independence of the $p$ random variables whose samples are given in the $p$ rows of the matrix $Z = W^T X$.

It is important to note that the integer $n$ is fixed by the chosen dataset $X$, while the integer $p$ defines the number of components the user wants to

compute. This number is, in most applications, smaller than $n$. Some contrast functions are able to deal with rectangular demixing matrices. Nevertheless, most of them are defined for square matrices only, such that the observations have to be preprocessed by means of prewhitening. Prewhitening is equivalent to Principal Component Analysis (PCA), which performs a Singular Value Decomposition of the matrix of the observations. The reduction of the dimension is achieved by retaining the dominant $p$-dimensional subspace, that is, the subspace related to the $p$ largest singular values. The matrix optimization is then applied on the prewhitened observations and identifies a square matrix of dimension $p$. Hence, the demixing matrix results from the product of two matrices. The first belongs to $\mathbb{R}^{n\times p}$ and is identified by prewhitening, while the second belongs to $\mathbb{R}^{p\times p}$ and results from the optimization of the contrast function. Most ICA algorithms present these two successive steps. They are denoted *prewhitening-based*.

Another important issue is to deal with the inherent symmetries of contrast functions. Symmetries are present because the measure of dependence between random variables must not be altered by scaling or by permutation of these variables. Optimizing a function with symmetries entails difficulties of theoretical (convergence analysis) and practical nature unless some constraints are introduced.

In the case of prewhitening-based ICA, where the whitenend matrix $X$ satisfies $XX^T = I$, it is common practice to restrict the matrix $W$ to be orthonormal, i.e., $W^TW = I$. This implies that the sample covariance matrix $ZZ^T$ is the identity matrix. Two options are conceivable to deal with the orthogonality constraint on $W$. First is to perform constrained optimization over a Euclidean space, i.e.,

$$\min_{W\in\mathbb{R}^{p\times p}} \gamma(W) \quad \text{such that} \quad W^TW = I.$$

This paper favors the second alternative which incorporates the constraints directly into the search space and performs unconstrained optimization over a nonlinear matrix manifold, i.e.,

$$\min_{W\in\mathcal{O}_p} \gamma(W) \quad \text{with } \mathcal{O}_p = \{W \in \mathbb{R}^{p\times p} | W^TW = I\}. \tag{2}$$

Most classical unconstrained optimization methods - such as gradient-descent, Newton, trust-region and conjugate gradient methods - have been generalized to the optimization over matrix manifolds, in particular over the orthogonal group $\mathcal{O}_p$. Developing efficient matrix algorithms that perform optimization on a matrix manifold is a topic of active research (see the monograph [AMS] and references therein).

## 3 Contrast functions

This section discusses the function $\gamma$ of the general problem statement (2). This function measures the statistical independence of random variables. It presents thus a global minimum at the solution of the ICA problem. Many contrast functions have been proposed in the literature, and we review some of them here.

### 3.1 Mutual information [Com94, LMI03]

Mutual information is a central notion of information theory [Mac02, CT06] that characterizes statistical independence. The mutual information $I(Z)$ of the multivariate random variable $Z = (z_1 \dots, z_p)$ is defined as the Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions,

$$I(Z) = \int p(z_1, \dots, z_p) \log \frac{p(z_1, \dots, z_p)}{p(z_1) \dots p(z_p)} dz_1 \dots dz_p.$$

The mutual information presents all the required properties for a contrast function: it is non negative and equals zero if and only if the variables $Z$ are statistically independent. Hence, its global minimum corresponds to the solution of the ICA problem.

Several approaches to compute efficiently an approximation to this quantity can be found in the literature. These approaches expand the mutual information in a sum of integrals that are expected to be more easily evaluated, namely the differential entropy and the negentropy. The differential entropy $S(z)$ and the negentropy $J(z)$ of a random variable $z$ are respectively defined by

$$S(z) = \int p(z) \log(p(z)) dz, \qquad \text{and} \qquad J(z) = S(g) - S(z),$$

where $g$ stands for a gaussian variable with same mean and variance as $z$. The mutual information can be expressed in terms of differential entropies as follows,

$$I(Z) = \sum_{i=1}^{p} S(z_i) - S(z_1, \dots, z_p).$$

A similar expansion in terms of negentropies in given by

$$I(Z) = J(z_1, \dots, z_p) - \sum_{i=1}^{p} J(z_i) + \frac{1}{2} \log \frac{\prod C_{ii}^Z}{|C^Z|},$$

where $C^Z$ denotes the covariance matrix of $Z$, $\prod C_{ii}^Z$ the product of its diagonal elements and $|\cdot|$ the determinant.

A contrast defined over the space of the demixing matrices is obtained once the demixing model $Z = W^T X$ is introduced within these two expansions, i.e,

$$\gamma(W) = \sum_{i=1}^{p} S(e_i^T W^T X) - \log(|W|) - S(x_1, \ldots, x_p), \tag{3}$$

and

$$\gamma(W) = J(x_1, \ldots, x_p) - \sum_{i=1}^{p} J(e_i^T W^T X) + \frac{1}{2} \log \frac{\prod C_{ii}^{W^T X}}{|C^{W^T X}|}, \tag{4}$$

where $e_i$ is the $i$th basis vector. More details about the derivation of these expressions can be found in [LMI03] for (3) and in [Com94] for (4). It should be noted that, once the observations are prewhitened and the demixing matrix is restricted to be orthogonal, both terms $\log(|W|)$ and $\frac{1}{2} \log \frac{\prod C_{ii}^{W^T X}}{|C^{W^T X}|}$ cancel.

At this point, statistical estimators are required to evaluate efficiently the differential entropy as well as the negentropy for a one-dimensional variable. Comon suggests using the Edgeworth expansion of a probability function in order to estimate the negentropy [Com94]. A truncated expansion up to statistics of fourth order leads to the following approximation,

$$J(z) \approx \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4,$$

where $\kappa_i$ denotes the cumulant of order $i$ of the standardized one-dimensional random variable $z$.

An efficient estimator of the differential entropy was derived by considering order statistics [LMI03]. Given a one-dimensional variable $z$ defined by its samples, the order statistics of $z$ is the set of samples $\{z^1, \ldots, z^N\}$ rearranged in non-decreasing order, i.e., $z^1 \leq \ldots \leq z^N$. The differential entropy of a one-dimensional variable $z$ defined by its order statistics $\{z^1, \ldots, z^N\}$ can be estimated by a simple formula,

$$\hat{S}(z) = \frac{1}{N-m} \sum_{j=1}^{N-m} \log \left( \frac{N+1}{m} (z^{(j+m)} - z^{(j)}) \right), \tag{5}$$

where $m$ is typically set to $\sqrt{N}$. This expression is derived from an estimator originally due to Vasicek [Vas76]. The contrast of the RADICAL algorithm [LMI03] is actually the function (3) where the differential entropies are evaluated with the estimator (5),

$$\gamma(W) = \sum_{i=1}^{p} \hat{S}(e_i^T W^T X) - \log(|W|) - S(x_1, \ldots, x_p). \tag{6}$$

### 3.2 $\mathcal{F}$-correlation [BJ03]

This contrast is based on a generalization of the Pearson correlation coefficient, called the $\mathcal{F}$-correlation. It is proven in [BJ03] that two random variables $z_1$ and $z_2$ are statistically independent if and only if the $\mathcal{F}$-correlation $\rho_{\mathcal{F}}$ vanishes, with $\rho_{\mathcal{F}}$ being defined by

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}(f_1(z_1) f_2(z_2)),$$

where $\operatorname{corr}(x, y)$ is the Pearson correlation coefficient between the random variables $x$ and $y$ and $\mathcal{F}$ is a vector space of functions from $\mathbb{R}$ to $\mathbb{R}$. A contrast for the two-dimensional ICA problem is thus given by

$$\gamma(W) = \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}(f_1(w_1^T X) f_2(w_2^T X)), \tag{7}$$

where $w_i$ is the $i$-th column of the matrix $W$. This quantity seems complex to evaluate since it involves an optimization over a space of infinite dimension. Nevertheless, the authors of [BJ03] showed that, by means of kernel methods [Sai88, SS01], this evaluation can be approximated by the solution of an eigenvalue problem of finite dimension. The $\mathcal{F}$-correlation $\rho_{\mathcal{F}}$ is estimated by the largest eigenvalue of a generalized eigenvalue problem of dimension $2N$ where $N$ stands for the number of samples.

The contrast function (7) allows the identification of only two independent components. The paper [BJ03] proposes a generalization to higher dimensions. The contrast remains the largest eigenvalue of a generalized eigenvalue problem, but of dimension $pN$, with $p$ being the number of components. This contrast is the core of the KernelICA algorithm [BJ03].

### 3.3 Non-gaussianity [HKO01]

Informally speaking, the central limit theorem states that the sum of independent random variables converges (in distribution) to a Gaussian variable as the number of terms tends to infinity. Thus, each linear combination of random variables is expected to be more gaussian than the original ones, which should be the most non-gaussian. A whole range of contrast functions is based on measuring the gaussianity (or non-gaussianity) of a one-dimensional random variable. The most intuitive expression for such a measure is given by

$$J(z) = (E[G(z)] - E[G(g)])^2, \tag{8}$$

where $E[\cdot]$ is the expectation operator, $G$ is a smooth function and $g$ is a gaussian variable with same mean and variance as $z$. A quadratic function $G$ enables to reveal features related to statistics up to the second order only,

whereas non-gaussianity involves higher order statistics. For this reason, $G$ should be non-quadratic. Some suggestions for the function $G$ are given in [HKO01]. For the particular choice of $G(z) = \frac{1}{4}z^4$, the distance to gaussianity becomes the square of the kurtosis $\kappa(z)$,

$$J(z) = \kappa(z)^2.$$

The kurtosis is a classical measure of non-gaussianity since it equals zero for gaussian variables and has a large absolute value for non-gaussian ones. The link between non-gaussianity and statistical independence is rigourously proven in the particular case of the kurtosis. A theorem in [Mat01] states that the kurtosis of the sum of two independent variables $z_1$ and $z_2$ presents a smaller absolute value than the largest absolute value of the kurtosis among these variables, i.e.,

$$|\kappa(z_1 + z_2)| \leq \max(|\kappa(z_1)|, |\kappa(z_2)|).$$

A contrast function is easily obtained from these measures of gaussianity by introducing the ICA model $z = w^T X$ in equation (8),

$$\gamma(w) = (E[G(w^T X)] - E[G(g)])^2, \tag{9}$$

where $w$ is a vector of $\mathbb{R}^n$. The maximization of this contrast results in the FastICA algorithm [HKO01], probably the most popular ICA algorithm. It is important to note that this contrast is one-unit based, i.e., it is defined for one column of the demixing matrix $W$. The notion of statistical independence becomes meaningful once the optimization is performed several times from different initial conditions and identifies different sources $z$.

### 3.4 Joint diagonalization of cumulant matrices [Car99]

Independent random variables are also characterized by the diagonality of some statistically motivated matrices. A necessary condition for statistically independent random variables is, for example, given by a diagonal covariance matrix. However, identifying the transformation $W$ that diagonalizes the covariance matrix of $Z = W^T X$ for some observations $X$ will only identify components that are uncorrelated but not independent.

In case of a zero-mean random variable, the covariance matrix can be considered as a particular case of the concept of *cumulant tensors*. The covariance matrix is then simply the cumulant tensor of second order. We refer to [LV04] for the definition and properties of these tensors. The essential issue for the present review is that cumulant tensors related to statistically independent variables are diagonal at any order. A particular matrix is derived from the fourth order cumulant tensor that presents the desired property of

being diagonal in case of independent variables. If $\mathcal{Q}_X^4$ denotes the fourth order cumulant tensor of the $p$-variate random variable $X$, the *cumulant matrix* $Q_X(M)$ related to a matrix $M$ is defined elementwise by

$$Q_X(M)|_{ij} = \sum_{k,l=1}^{p} \mathcal{Q}_X^4|_{ijkl} M_{kl},$$

where $\mathcal{Q}_X^4|_{ijkl}$ denotes the element at position $(i, j, k, l)$ of the fourth order cumulant tensor. The cumulant matrix can be efficiently evaluated without the computation of the whole tensor $\mathcal{Q}_X^4$ thanks to the following expression,

$$\begin{aligned} Q_X(M) = E[(X^T M X) X X^T] \\ - E[X X^T]\mathrm{tr}(M E[X X^T]) - E[X X^T](M + M^T)E[X X^T], \end{aligned}$$

where $\mathrm{tr}(\cdot)$ denotes the trace. This expression is valid only if the variable $X$ has a zero mean.

An important property of the cumulant matrix is that the orthogonal transform $Z = W^T X$ results in a similarity transform for $Q(M)$, i.e.,

$$Q_Z(M) = W^T Q_X(M) W$$

whatever the matrix $M$.

A set of matrices that are simultaneously diagonal can be constructed by picking several matrices $M$. The maximal set of cumulant matrices is obtained whenever the matrices $M$ form an orthogonal basis for the linear space of $p \times p$ symmetric matrices. The matrix $W$ that diagonalizes simultaneously all the cumulant matrices performs a projection of the observations $X$ on random variables $Z$ that are statistically independent. However, it is usually impossible to identify a joint diagonalizer $W$ that diagonalizes exactly all these matrices. This suggests defining ICA algorithms that optimize the joint diagonalization of them.

Several cost functions for this approximate joint diagonalization are conceivable, but a frequently encountered one is the following,

$$\gamma(W) = \sum_i \|\mathrm{off}(W^T Q_X(M_i) W)\|_F^2, \tag{10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathrm{off}(A)$ is a matrix with entries identical to those of $A$ except on the diagonal, which contains only zero-valued elements. Optimizing (10) means that one minimizes the sum of the squares of all non-diagonal elements of the cumulant matrices. This is consistent with performing the best approximate joint diagonalization of these matrices. Thanks to the diagonal property of the cumulant matrices in case

of independent variables, this cost function is at the same time a contrast for ICA.

This contrast is used by the famous JADE algorithm [Car99]. We mention that other type of matrices have been defined in the literature that are also diagonal in the case of independent random variables. The SOBI algorithm (Second Order Based Identification), for example, performs ICA by the approximate joint diagonalization of matrices that involve only second order statistics [BAMCM97].

## 4 Matrix manifolds for ICA

Each contrast function considered in the previous section possesses two symmetries. First, since the statistical independence of random variables is not affected by a scaling, the following equality holds,

$$\gamma(W) = \gamma(W\Lambda),$$

whenever $\Lambda$ is a diagonal matrix. Likewise, the permutation of random variables does not affect their statistical independence, i.e.,

$$\gamma(W) = \gamma(WP),$$

where $P$ is a permutation matrix. Because of these symmetries, the global minimum of the contrast is not a single point but a set of subspaces of $\mathbb{R}^{p\times p}$. Optimizing cost functions that present continuous symmetries is a difficult task. Therefore, constraints have to be added to restrict these subspaces to a set of discrete points. In the present case, the scaling symmetry disappears if each column of $W$ is set to a fixed norm, for example to a unit-norm. This constraint set defines the so-called oblique manifold [AG06],

$$\mathcal{OB}_p = \{W \in \mathrm{GL}(p) : \mathrm{diag}(W^TW) = I\},$$

where $\mathrm{GL}(p)$ is the set of all invertible $p \times p$ matrices. This manifold gets rid of the continuous scaling symmetry. The permutation symmetry defines the global optimum as a set of discrete points. This is suitable to most optimization algorithms. Hence, the oblique manifold is the most general manifold to perform ICA in an efficient way.

Usually, some further constraints are imposed. Prewhitening-based ICA algorithms preprocess the data by means of Principal Component Analysis (PCA). It is shown in [Com94] that this allows us to restrict the ICA optimization to the orthogonal group,

$$\mathcal{O}_p = \{W \in \mathbb{R}^{p\times p} | W^TW = I_p\}.$$

Furthermore, this manifold ensures a good conditioning of the optimization algorithms.

As mentioned in the previous section, the FastICA contrast is defined for one column of the demixing matrix $W$, while most of the other contrasts are defined on a matrix space. In order to remove the scaling symmetry, the optimization of that contrast is performed on the sphere,

$$S^{n-1} = \{w \in \mathbb{R}^n | w^T w = 1\}.$$

Some implementations of the FastICA algorithm perform in parallel several optimizations of the contrast (9) starting from different initial conditions [HKO01]. In order to prevent convergence toward identical minima, the parallel iterates are reorthogonalized at each iteration, so that the algorithm simultaneously identifies $p$ independent columns of the demixing matrix. Since these columns are orthogonal, the data must be preprocessed by PCA, but the square constraint $p = n$ disappears. These implementations are equivalent to an optimization over the Stiefel manifold,

$$\mathrm{St}(n, p) = \{W \in \mathbb{R}^{n \times p} | W^T W = I_p\}.$$

## 5 Optimization algorithms

### 5.1 Line-search algorithms

Many optimization algorithms on a Euclidean space are based on the following update formula,

$$x_{k+1} = x_k + t_k \eta_k. \tag{11}$$

which consists to move from the current iterate $x_k$ in the search direction $\eta_k$ with a certain step size $t_k$ to identify the next iterate $x_{k+1}$. The search direction and the step size are chosen such that the cost function decreases sufficiently at each iteration. The search direction is usually set to the opposite of the gradient of the cost function $\gamma$ at the current iterate, i.e.,

$$\eta_k = -\mathrm{grad}\gamma(x_k).$$

The iterate is thus moving in the direction of steepest-descent. The step size has then to be selected in order to induce a significant decrease of the cost function. Iteration (11) is, however, valid only in case the iterates belong to a Euclidean space.

On non-Euclidean manifolds, the update formula (11) is generalized to

$$W_{k+1} = R_{W_k}(t_k \eta_k),$$

where $W_k$ and $W_{k+1}$ are two successive iterates on the manifold $\mathcal{M}$, $t_k$ is a scalar and $\eta_k$ belongs to $T_{W_k}\mathcal{M}$, the tangent space to $\mathcal{M}$ at $W_k$. The retraction $R_W(\eta)$ is a mapping from the tangent plane to the manifold. More details about this concept can be found in [AMS].

The search direction $\eta_k$ is, as above, set to the opposite of the gradient of the cost function $\gamma$ at $W_k$,

$$\eta_k = -\mathrm{grad}\gamma(W_k).$$

In case of a manifold $\mathcal{M}$ embedded in a Euclidean space $\bar{\mathcal{M}}$, the gradient at $W \in \mathcal{M}$ is simply computed as the orthogonal projection onto $T_W\mathcal{M}$ of the gradient in the embedding space, i.e.,

$$\mathrm{grad}\gamma(W) = \mathrm{P}_W \mathrm{grad}\bar{\gamma}(W),$$

where $\bar{\gamma}$ denotes a smooth extension on $\bar{\mathcal{M}}$ of the cost function $\gamma$, i.e.,

$$\bar{\gamma}(W) = \gamma(W), \quad \forall\, W \in \mathcal{M}.$$

We refer to [AMS] for more details about the orthogonal projector $\mathrm{P}_W$.

The gradient in the embedding Euclidean space is defined, as usual, from the directional derivative,

$$\langle \mathrm{grad}\bar{\gamma}(W), \zeta \rangle = D\bar{\gamma}(W)[\zeta] = \lim_{t \to 0} \frac{\bar{\gamma}(W + t\zeta) - \bar{\gamma}(W)}{t},$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product.

To complete the description of the gradient-descent algorithm, the choice of the step size $t_k$ has to be discussed. Several alternatives are conceivable. First, an exact line-search identifies the minimum in the direction of search. Such an optimization is usually tricky and requires a huge computational effort. A good alternative is given by the Armijo step size $t^A$. This step size is defined by

$$t^A = \beta^m \alpha,$$

with the scalars $\alpha > 0$, $\beta \in (0, 1)$ and $m$ being the first nonnegative integer such that,

$$\gamma(W) - \gamma(R_W(\beta^m \alpha)) \geq -\sigma \langle \mathrm{grad}\gamma(W), \beta^m \alpha \rangle_W,$$

where $W$ is the current iterate on $\mathcal{M}$, $\sigma \in (0, 1)$, $R_W(\eta)$ is a retraction.

The KernelICA algorithm performs a gradient-descent optimization of the $\mathcal{F}$-correlation $\rho_{\mathcal{F}}$ on the orthogonal group $\mathcal{O}_p$ [BJ03]. The projection operator and the retraction are respectively given by

$$P_W(\eta) = (I - WW^T)\eta \qquad \text{and} \qquad R_W(\eta) = W\exp(W^T\eta).$$

We refer to the original paper [BJ03] for the details about the derivation of the gradient.

A gradient-descent algorithm for the minimization over the orthogonal group of the contrast function (6), originally dedicated to the RADICAL algorithm [LMI03], was recently proposed by us in [JTAS07].

Not only line-search methods generalize on nonlinear manifolds. We mention, among others, trust-regions and conjugate gradient algorithms. More details about these methods can be found in [AMS].

### 5.2 FastICA

FastICA algorithms perform the optimization of the cost function $\gamma$ in one direction at the time,

$$\max_{w\in\mathbb{R}^n} \gamma(w), \quad \text{such that} \quad w^T w = 1, \tag{12}$$

where $w$ is one column of the demixing matrix $W$. In this case, the orthonormal constraint $W^TW = I_n$ reduces to a spherical constraint $w^Tw = 1$.

The FastICA approach then exploits standard constrained optimization schemes. The solution of the problem (12) has to satisfy the Kuhn-Tucker condition

$$\frac{\partial\gamma(w)}{\partial w} - \beta w = 0,$$

where $w \in \mathbb{R}^n$ and $\beta$ is a Lagrange multiplier. A Newton method to solve this equation results in the iteration,

$$w^+ = w - \left(\frac{\partial^2\gamma(w)}{\partial w^2} - \beta I\right)^{-1} \cdot \left(\frac{\partial\gamma(w)}{\partial w} - \beta w\right), \tag{13}$$

where $w^+$ denotes the new iterate. The central point of the FastICA algorithm is to approximate the matrix inversion of the Hessian by a simple scalar inversion. It is shown in [HKO01] that once the data is prewhitened, the Hessian of the contrast is close to a scalar matrix,

$$\frac{\partial^2\gamma(w)}{\partial w^2} \approx \frac{\partial^2\tilde{\gamma}(z)}{\partial z^2} I,$$

with

$$\tilde{\gamma} : \mathbb{R} \to \mathbb{R} \quad \text{such that} \quad \tilde{\gamma}(w^TX) = \gamma(w),$$

where $X$ is the data. Hence, iteration (13) can be approximated by

$$w^+ = w - \frac{1}{\frac{\partial^2 \tilde{\gamma}(z)}{\partial z^2} - \beta} \cdot \left( \frac{\partial \gamma(w)}{\partial w} - \beta w \right). \tag{14}$$

Multiplying both sides of (14) by $(\beta - \frac{\partial^2 \tilde{\gamma}(z)}{\partial z^2})$ results in,

$$w^+ = \left( \beta - \frac{\partial^2 \tilde{\gamma}(z)}{\partial z^2} \right) \cdot w + \left( \frac{\partial \gamma(w)}{\partial w} - \beta w \right) = \frac{\partial \gamma(w)}{\partial w} - \frac{\partial^2 \tilde{\gamma}(z)}{\partial z^2} w.$$

The new iterate is normalized at each iteration to a unit-norm to ensure the stability of the algorithm,

$$w^+ = \frac{w^+}{\|w^+\|}.$$

Hence, the FastICA algorithm consists in repeating the iteration

$$\begin{cases} w^+ = \frac{\partial \gamma(w)}{\partial w} - \frac{\partial^2 \tilde{\gamma}(z)}{\partial (z)^2} w \ , \quad \text{with } z = w^T X, \\ w^+ = \frac{w^+}{\|w^+\|}. \end{cases} \tag{15}$$

More details about this algorithm can be found in [HKO01].

The algorithm (15) identifies one column only of the demixing matrix $W$. Nevertheless, it can be used to reconstruct several columns of that matrix by means of a deflation technique. Assuming prewhitening of the data, the demixing matrix is orthogonal. Suppose that $p$ columns $\{w_1, \ldots, w_p\}$ of the whole matrix $W$ have been computed. The one-unit algorithm will converge to a new vector $w_{p+1}$ that is orthogonal to the already known directions if, after each iteration, the projections of these directions are subtracted,

$$w_{p+1}^+ = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T w_j w_j.$$

The drawback of any deflation scheme is that small computational errors are amplified by the computation of new vectors. Thus, the first computed directions should be accurately estimated. The last ones should be expected to be tagged with larger errors.
An alternative is the symmetric orthogonalization, which performs all the computations in parallel without favoring some directions. Each one-unit algorithm is randomly initialized and the iterates are reorthogonalized after each iteration according to,

$$W^+ = (WW^T)^{-\frac{1}{2}} W.$$

The matrix square root can be avoided by means of an iterative algorithm that is described in [HKO01].

### 5.3 Jacobi rotations

Jacobi rotations provide a classical optimization method on the orthogonal group [GVL96, Com94, Car99]. The iterates are constrained on the orthogonal group by successive multiplication by special orthogonal matrices $W_k$ containing a single parameter,

$$W_k = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos(\alpha) & \cdots & \sin(\alpha) & & \\ & & & \ddots & & & \\ & & \vdots & & \vdots & & \\ & & & & & \ddots & \\ & & -\sin(\alpha) & \cdots & \cos(\alpha) & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ {\scriptstyle (i)} \\ \\ \\ \\ {\scriptstyle (j)} \\ \\ \end{matrix} \begin{matrix} \\ \\ {\scriptstyle (i)} \\ \\ \\ \\ {\scriptstyle (j)} \\ \\ \end{matrix} .$$

Such matrices achieve a planar rotation of angle $\alpha$ in the subspace spanned by the directions $(i)$ and $(j)$. At each iteration, a new subspace is selected and the best rotation is computed to maximize the cost function. Hence, only a one-dimensional optimization problem has to be solved at each iteration. In case of the JADE algorithm [Car99], this task is performed analytically such that an explicit expression of the optimal angle $\alpha$ is available. The RADICAL algorithm on the other hand [LMI03] performs the global minimization over that parameter by exhaustive search on $[0, 2\pi]$, or more precisely on $[0, \frac{\pi}{2}]$ because of the permutation symmetry of the contrast function.

## 6 Analysis of gene expression data by ICA

### 6.1 Some issues about the application of ICA

To fix the notations, let us define the gene expression matrix $X$ such that its element $(i, j)$ corresponds to the expression level of gene $i$ in the $j$th experiment. $X$ is thus a $n \times N$ matrix, where $n$ is the number of analyzed genes and $N$ is the number of experiments. Note that $n$ is usually much larger than $N$. The breast cancer database considered for the following analysis is related to $n = 17816$ genes and $N = 286$ patients. The number $n$ is typical for genome-wide expression data while $N = 286$ is fairly large in comparison with most other profiling studies in breast cancer.

Two modelling hypothesis underlie the application of ICA to gene expression data. First is that the expression level of a gene is a linear superposition of biological processes, some of which try to express it, while others try to repress it. Specifically, ICA performs an approximate decomposition of the gene expression matrix into two smaller matrices $A$ and $B$ that are respectively $n \times p$ and $p \times N$ with $p < N$, i.e.,

$$X \approx AB. \tag{16}$$

Figure 1 provides a clearer idea of this decomposition by representing matrices with Hinton diagrams. In such diagrams, each value is displayed by a square whose size is an image of the magnitude.

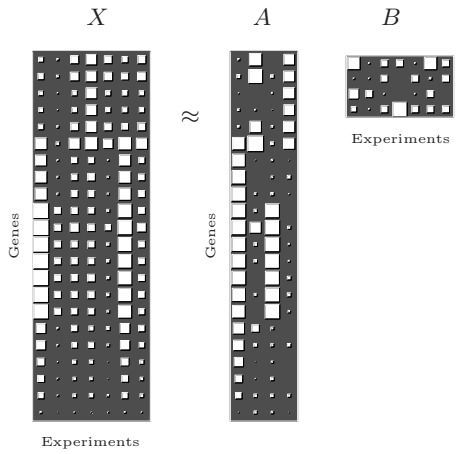


**Fig. 1.** ICA decomposition of a gene expression database $X$.

Each column of $A$ is a vector in the gene space that is called an *expression mode*. In the same spirit, each row of $B$ is a vector in the samples space that stands for the *activity* of the corresponding expression mode. This means that the expression of one particular gene across a set of samples (e.g. different patients) is modelled as the superposition of a restricted number of expression modes (typically about 10). The activity is the relative weight of an expression mode across the samples. The matrices $A$ and $B$ are selected to minimize the error between $X$ and $AB$ in the least-square sense. The second hypothesis imposes a notion of independence somewhere in the underlying biology. Hence, the matrices $A$ and $B$ have to maximize a certain measure of statistical independence.

Two alternatives are conceivable for the statistical independence assumption within the model (16). Either the independence is set in sample space, i.e., the rows of $B$ stand for independent random variables (model I), or it is

set in gene space, i.e., the columns of $A$ stand for independent random variables (model II). We next describe the underlying and distinct motivations for the two models.

In the case of model I, the algorithm seeks independent components across samples, which amounts to finding non-gaussian distributions for the rows of $B$. It is expected that these non-gaussian projections separate samples into biologically relevant subtypes, although there is no requirement that this be so. As such, the algorithm performs a kind of unsupervised projection pursuit.

In contrast model II seeks independent components across genes. There is a strong biological motivation for this model, since it is most natural to assume that the measured gene expression levels represent the net effect of many independent biological processes, which may or may not have different activity levels across samples. Furthermore, it is natural to assume that most genes in an expression mode do not play an active role in it, and that it is only a small percentage of genes which are relevant to any given expression mode. This is tantamount to assuming that the distribution of weights in the columns of $A$ are supergaussian (leptokurtic), which fits in well with the ICA model II.

An issue of importance deals with the estimation of $p$, the number of independent components that underlie the gene expression data set. Some approaches compute the probability of observing the data $X$ if the underlying model contains $p$ components. This enables a rough estimation of the number of independent components. The Bayesian Information Criterion (BIC) in a maximum likelihood framework [HLK01] or using the evidence bound in a variational Bayesian approach [Min00] are examples of such methods. In the present study, we choose a fixed number of components for all algorithms. The correct estimation of the number of components seems difficult because of the small number $N$ of samples available.

### 6.2 Evaluation of the biological relevance of the expression modes

The general objective of PCA and ICA methods is to identify a small set of variables, in terms of which the data can be more easily interpreted. Previous applications of ICA to gene expression data have evaluated the results by means of the Gene Ontology (GO) framework. However, this does not provide the best framework in which to validate these methods, since genes with the same GO-annotation may not necessarily be part of the same biological pathway, and vice versa, genes that are part of the same pathway may have quite distinct biological functions. Instead of GO, we propose to use the framework of biological pathways, since it is the alteration pattern of specific pathways that underlies the cancer-cell phenotype. The independent components derived from ICA are expected to summarize the net effect of

independent altered transcriptional programs in cancer, and as such, should map closer to aberrations in existing biological and cancer-related pathways. While research in cancer biology is still at the stage of trying to elucidate all the pathways that may be involved, several efforts are underway in building up pathway databases. Some of these pathways have been curated from various sources, while others were determined by specific experiments. Each pathway in these databases is essentially a list of genes that are known to participate together when a certain biological function is required. In this work, we evaluate the PCA and ICA methods against their ability to correlate expression modes with known pathways. To our knowledge, the PCA and ICA methods were never evaluated in the explicit context of biological pathways.

Let us specify more concretely the concept of mapping between a pathway and an expression mode. Each expression mode is a list of all the genes with an associated weight. Since a pathway is just a list of some specific genes that are known to present a linked activity, the expression modes require some post-processing that consists in selecting the genes with major weights. To identify the genes that are differentially activated, it is common to impose a threshold of typically 2 or 3 standard deviations from the mean of the distribution of the inferred weights. A gene list that looks like a pathway is thus obtained by selecting the genes with an absolute weight that exceeds this threshold. A stringent threshold at three standard deviations was chosen in the present study to reveal the most relevant pathways captured by each of the expression modes.

If the application of ICA/PCA in the gene expression context is biologically well-founded, each expression mode should be strongly tagged by a specific pathway or by a superposition of some distinct pathways that are highly dependent. A quantitative measure of the enrichment level of a given expression mode $i$ in a given pathway $p$ has to be defined. Let $n_i$ denote the number of genes in the expression mode and $n_p$ denote the number of genes in the pathway. Further, let $d_i$ denote the number of genes selected in the expression mode $i$ and $t_{ip}$ the number of genes from pathway $p$ among the selected $d_i$ genes. Under the null-hypothesis, where the selected genes are chosen randomly, the number $t_{ip}$ follows a hypergeometric distribution [BS04]. Specifically, the probability distribution is

$$\begin{aligned} P(t) &= \binom{d_i}{t} \prod_{j=0}^{t-1} \frac{n_p - j}{n_i - j} \prod_{j=0}^{j=d_i-t-1} \frac{n_i - n_p - j}{n_i - t - j} \\ &= \frac{\binom{n_p}{t}\binom{n_i - n_p}{d_i - t}}{\binom{n_i}{d_i}}. \end{aligned}$$

A probability can thus be computed as $P(t > t_{ip})$. This quantity enables to estimate for each mode-pathway pair how the mode is enriched in terms of genes from that particular pathway. This mode-pathway association will be said significant if the P-value $P(t > t_{ip})$ is less than a certain threshold. To evaluate the biological significance of the ICA approach, a large set of pathways is selected and a pathway enrichment index (PEI) is defined as the fraction of biological pathways that are found to be enriched in at least one expression mode.

### 6.3 Results obtained on the breast cancer microarray data set

PCA as well as the four ICA methods detailed in the first sections of this paper (i.e., JADE [Car99], RADICAL [Lie02], KernelICA [BJ03] and FastICA [HKO01]) have been applied to one of the largest breast cancer microarray data set available [WKZ+05]. The analysis was performed for both models I and II. The number of components $p$ was fixed to 10 in each study. Since the four ICA algorithms are all prewhitening-based, the reduction of the dimensions of the problem from $N$, the number of experiments, to $p$ is simply done by Singular Value Decomposition (SVD) during the prewhitening step. The ICA step computes thereafter a square demixing matrix of dimensions $p \times p$.

To evaluate the biological significance of the results, we compiled a list of 536 pathways that are known to be directly or indirectly involved in cancer biology. 522 of these pathways come from the Molecular Signature Database MSigDB [STM+05]. The others are known oncogenic pathways recently derived in [BYC+05] and cancer-signalling pathways coming from the resource NETPATH (*www.netpath.org*). Figure 2 shows the pathway enrichment index (PEI) based on these 536 pathways for the five methods for both models I and II.

The same kind of analysis was performed on the reduced set of the oncogenic pathways and the cancer-signalling pathways of NETPATH. Since these 14 pathways are frequently altered in cancer, many of them are expected to be captured by the expression modes. The PEI related to them are illustrated on Figure 3.

Both Figures 2 and 3 indicate that ICA achieves a more realistic representation of the gene expression data than PCA. Furthermore, the PEI values are clearly higher for model II than for model I. Hence, the ICA-model with the independence assumption stated in the gene space seems to be the most efficient approach to unveil the biological significance of the gene expression data. It is however difficult to discriminate between different ICA algorithms.

The present work is part of a larger project that investigates the assets of the ICA approach for the biological interpretation of microarray databases.

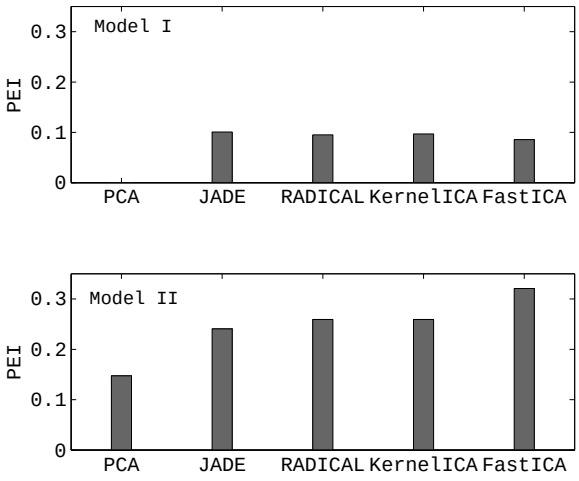


**Fig. 2.** PEI based on a set of 536 pathways for both models I and II.

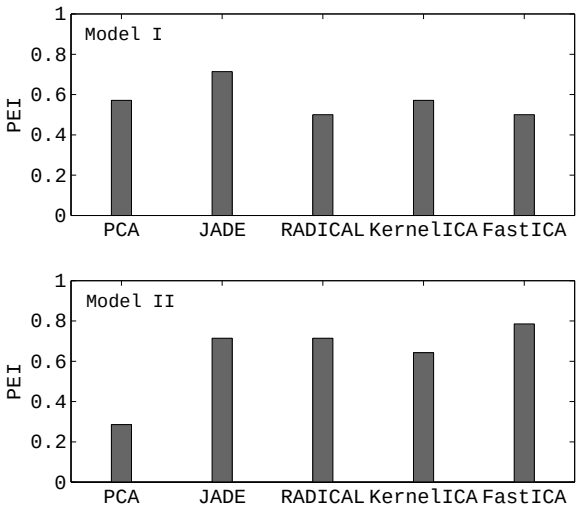


**Fig. 3.** PEI based on a set of 14 cancer-signalling and oncogenic pathways for both models I and II.

A deeper analysis of the ICA decomposition has been performed on a total of nine microarray data sets related to breast cancer, gastric cancer and lymphoma. This large study also favors the use of ICA with model II for the analysis of gene expression data. It highlights that ICA algorithms are able

to extract biological information not detected by PCA methods. More details about this study can be found in [TJA+07].

## 7 Conclusion

DNA microarrays enable new perspectives in biomedical research and especially in the understanding of some cellular mechanisms. This is of outmost interest for the treatment of cancer diseases. This emerging technology provides such a huge amount of data that unsupervised algorithms are required to automatically unveil the biological processes that have led to the observed transcriptome. The present paper reviews some standard algorithms to perform Independent Component Analysis and emphases their common feature, namely the optimization of a measure of statistical independence (the contrast) over a matrix manifold. The paper then illustrates the way to use these algorithms in the context of gene expression data. Even if the application of ICA to gene expression data sets is not a new idea, the evaluation of the results in the explicit context of biological pathways, has never been performed before. The main conclusion of this study is the significant outperformance of the ICA approach against Principal Component Analysis (PCA). The ICA model, with the statistical independence assumption stated in the gene space, seems to be a realistic representation of the mechanisms that determine the gene expression levels. ICA shows significant promise for the analysis of DNA microarray databases.

## Acknowledgement

This work was supported by the Belgian National Fund for Scientific Research (FNRS) through a Research Fellowship at the University of Liège (MJ), by Microsoft Research through a Research Fellowship at Peterhouse, Cambridge (PAA), by a grant from Cancer Research UK and by a grant from the Isaac Newon Trust to Simon Tavare (AET). This paper presents research results of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

## References

[ABB03] O. Alter, P. O. Brown, and D. Botstein, *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.*, Proc Natl Acad Sci USA **100** (2003), no. 6, 3351–3356.

[AG06] P.-A. Absil and K.A. Gallivan, *Joint diagonalization on the oblique manifold for independent component analysis*, Proceedings of ICASSP2006, 2006.

[AMS] P.A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, To appear.

[BAMCM97] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, *A blind source separation technique using second-order statistics*, IEEE Transactions on Signal Processing **45** (1997), 434–444.

[BJ03] F. R. Bach and M. I. Jordan, *Kernel independent component analysis*, Journal of Machine Learning Research **3** (2003), 1–48.

[BS04] Tim Beißbarth and Terence P. Speed, *GOstat: find statistically over-represented gene ontologies within a group of genes*, Bioinformatics **20** (2004), no. 9, 1464–1465.

[BYC+05] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, Jeffrey R. Marks, H. K. Dressman, M. West, and J. R. Nevins, *Oncogenic pathway signatures in human cancers as a guide to targeted therapies*, Nature (2005).

[Car99] J.-F. Cardoso, *High-order contrasts for independent component analysis*, Neural Computation **11** (1999), no. 1, 157–192.

[Com94] P. Comon, *Independent Component Analysis, a new concept ?*, Signal Processing, Elsevier **36** (1994), no. 3, 287–314, Special issue on Higher-Order Statistics.

[CT06] T. M. Cover and J. A. Thomas, *Elements of information theory (wiley series in telecommunications and signal processing)*, Wiley-Interscience, 2006.

[GVL96] G. H. Golub and C. F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, October 1996.

[HKO01] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, 2001.

[HLK01] L. Hansen, J. Larsen, and T. Kolenda, *Blind detection of independent dynamic components*, 2001.

[JTAS07] M. Journée, A. E. Teschendorff, P.-A. Absil, and R. Sepulchre, *Geometric optimization methods for independent component analysis applied on gene expression data*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), April 2007.

[LB03] S.-I. Lee and S. Batzoglou, *Application of independent component analysis to microarrays*, Genome Biology **4** (2003), R76.

[Lie02] W. Liebermeister, *Linear modes of gene expression determined by independent component analysis*, Bioinformatics **18** (2002), 51–60.

[LMI03] E.G. Learned-Miller and J. W. Fisher III, *ICA using spacings estimates of entropy*, Journal of Machine Learning Research **4** (2003), 1271–1295.

[LV04] L. De Lathauwer and J. Vandewalle, *Dimensionality reduction in higher-order signal processing and rank-*$(R_1, R_2, ..., R_n)$ *reduction in multilinear algebra*, Lin. Alg. Appl. **391** (2004), 31–55.

[Mac02] D. J. C. Mackay, *Information theory, inference & learning algorithms*, Cambridge University Press, June 2002.

[Mat01] H. Mathis, *Nonlinear functions for blind separation and equalization*, Ph.D. thesis, Swiss Federal Institute of Technology, Zrich, Switzerland, 2001.

[Min00] T. P. Minka, *Automatic choice of dimensionality for PCA*, NIPS, 2000, pp. 598–604.

[MMSM02] A.-M. Martoglio, J. W. Miskin, S. K. Smith, and D. J. C. MacKay, *A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer*, Bioinformatics **18** (2002), no. 12, 1617–1624.

[RCTH05] A. Riva, A.-S. Carpentier, B. Torrésani, and A. Hénaut, *Comments on selected fundamental aspects of microarray analysis*, Computational Biology and Chemistry **29** (2005), no. 5, 319–336.

[Sai88] S. Saitoh, *Theory of reproducing kernels and its applications*, Longman Scientific & Technical, Harlow, England, 1988.

[SHK+03] S. A. Saidi, C. M. Holland, D. P. Kreil, D. J. C. MacKay, D. S. Charnock-Jones, C. G. Print, and S. K. Smith, *Independent component analysis of microarray data in the study of endometrial cancer*, Oncogene **23** (2003), no. 39, 6677–6683.

[SS01] B. Scholkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, MA, USA, 2001.

[STM+05] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, *From the cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS **102** (2005), no. 43, 15545–15550.

[TJA+07] A. E. Teschendorff, M Journée, P.-A. Absil, R. Sepulchre, and C. Caldas, *Elucidating the altered transcriptional programs in breast cancer using independent component analysis*, submitted to PLoS Biology (2007).

[Vas76] O. Vasicek, *A test for normality based on sample entropy*, Journal of the Royal Statistical Society: Series B **38** (1976), 54–59.

[WKZ+05] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*, Lancet **365** (2005), no. 9460, 671–679.