Estimating the proportions in a mixed sample using transcriptomics

Bertrand Clarke\textsuperscript{a,} and Jennifer Clarke\textsuperscript{b}

Received 11 August 2014; Accepted 23 September 2014

Often, biomedical researchers take a tissue sample from a tumour that unintentionally contains tumour cells and stromal cells so that gene expression from the sample represents two distinct kinds of cells. This is tolerated because techniques for separating tumour and stromal cells are expensive and time-consuming and the act of separation can alter gene expression. So, it is desirable to have a technique for estimating the proportion of tumour cells in a mixed sample to improve detection of differential expression in cancer cells. 2014 The Authors. Stat Published by John Wiley & Sons Ltd.

Keywords: deconvolution; gene expression; mixed sample; proportion

1 Introduction

Metagenomics is defined as the study of multiple genomes either from different species or from the same species but different states. It has become an important focus for researchers in several biomedical fields. The reason is that outside unusually well-controlled experiments, genetic material or measurements usually represent two or more distinct biological sources, for example, cells, tissue types, or organisms. Accordingly, methods for analysing metagenomic samples that take into account multiple sources must be developed.

Specifically, physical samples containing two or more kinds of tissue are frequently obtained from a subject even though a single tissue type is sought. When this happens, and the sample is used to generate gene expression data, it is unclear how much of the expression is due to the tissue of interest and how much is due to the contaminating tissue. This is common in cancer research: the part of the tissue sample of interest is often the tumour; however, the tissue also contains stromal cells. So, when a whole genome expression array is carried out for such a sample, we must correct for the presence of stromal cells to permit determination of whether the expression of a gene is significantly higher or lower in the cancer cells as compared with stromal cells or with a separate normal reference. This problem is often referred to as computational deconvolution or purification (Quon et al., 2013). We assume that expression values from a second sample of pure stromal cells are available, and we want to estimate the proportion $p$ of stromal cells in the mixed sample. It will be seen that one way to do this is to transform the raw data, thereby introducing a parameter $\alpha$ and using the transformed data to estimate $\rho$. This will necessitate estimating $\alpha$ as well.

\textsuperscript{a}Department of Statistics, University of Nebraska–Lincoln, Lincoln, NE 68583, USA
\textsuperscript{b}Department of Statistics and the Department of Food Science and Technology, University of Nebraska–Lincoln, Lincoln, NE 68583, USA

*Email: bclarke3@unl.edu

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Our procedure for estimating \( \alpha \) and \( p \) was presented by Clarke et al. (2010) where details on the implementation and evidence that it works well in practice in numerous and diverse contexts can be found. The estimators are based on ratios (Gosink et al., 2007). The estimator for the transformation parameter comes from graphing the mean of a ratio of gene expressions as a function of a multiplicative parameter in a logarithmic transform and looking for the minimum radius of curvature. The estimator for the proportion is based on minimizing a similar ratio.

Here, we present a theoretical justification for our approach to computational deconvolution; to the best of our knowledge, no other technique for estimating \( p \) has a formal consistency proof. Other approaches to deconvolution are discussed by Quon et al. (2013) and Ahn et al. (2013); hence, we mention them briefly here. One is due to Lu et al. (2003). The idea is that genes known to be expressed exclusively in one cell type can be used to estimate the proportion of expression coming from that type. This method depends on known tissue-specific or cell-specific genes and technology that can detect their expression with little or no cross-hybridization. A second approach requires expression profiles from each cell type. For instance, Wang et al. (2006) use a method similar to that of Lu et al. (2003) to generate estimates by obtaining solutions to linear equations via simulated annealing. Alternatively, Quon & Morris (2009) use a latent Dirichlet allocation model with [estimation via the expectation-maximization algorithm] in place of the linear equations and simulated annealing. These methods are not applicable in our context because we do not assume pure expression profiles.

There are other methods for expression deconvolution that do not require estimation of \( p \) (or \( \alpha \)). Several researchers have used expression data from purified reference tissue types to determine the expression of each tissue type in heterologous samples (Venet et al., 2004; Lahdesmaki et al., 2005). Another method uses proportions of each sample or cell type, assessed by pathologists, to establish tissue-specific expression. Stuart et al. (2004) use linear regression models, regressing expression on fractional content of each cell type, to estimate the expected cell type expression as the regression coefficients. Shen-Orr et al. (2010) extended this method to detect differential expression. Ghosh (2004) developed a method to determine differential expression in the presence of mixed cell populations in which a pathologist's assessments of the proportions of each cell type were used in a hierarchical mixture model. More recently, Quon et al. (2013) used sets of tumour and healthy tissue expression profiles to generate a purified cancer profile for each tumour sample, and an estimate of the proportion of RNA originating from cancerous cells. Finally, Ahn et al. (2013) developed an approach that estimates both tissue proportions and tissue-specific expressions using raw measured data from groups of samples. However, none of these approaches have a formal theory.

The structure of this paper is as follows. In Section 2, we formulate a model in which both \( p \) and \( \alpha \) are well defined as population quantities and provide estimators for them. In Section 3, we establish consistency of these estimators. In Section 4, we discuss several issues related to the proof and the problem.

# Defining the setting

The gene expression from two samples – one mixed tumour and stroma and the other pure stroma – can be regarded as a collection of bivariate two-stage experiments. The first stage is the selection of a gene from a population of genes, and the second stage is the measurement of the expression of the gene in the stromal sample and the mixed sample. Thus, we have a distribution of the form

\[
P(U^*_j, V^*_j \mid J = j)
\]

in which \( U^*_j \) is the gene expression from the pure stromal sample for gene \( j = 1, \ldots, m \), while \( V^*_j \) is the gene expression from the mixed sample. Note that \( J \) is regarded as a random variable, so if the subjects are independent and identical,
(1) holds for each of them. Later, \( m \) will be allowed to increase slowly as a mathematical convenience for proofs. In fact, \( m \) is usually large, so limits over \( m \) are realistic.

Let us suppose that the gene expression from the mixed sample can be expressed in terms of the gene expression from samples of pure stroma and pure tumour. So, for each \( j \),

\[
V_j^* = pU_j^* + (1-p)W_j^*
\]

where \( p \in [0, 1] \) is the proportion of stromal tissue expression in the sample and \( W_j^* \) is what the expression for the \( j \)-th gene in a pure tumour sample would be. It is understood that \( V_j^* \), \( U_j^* \), and \( W_j^* \) are all taken per unit tissue. In (2), we know \( V_j^* \) and \( U_j^* \), and we want to find \( p \) so we can determine \( W_j^* \). (An extension to three or more components is possible; see Remark 3 in Section 4.)

Heuristically, we can look at \( V_j^*/U_j^* = p + (1-p)(W_j^*/U_j^*) \) and minimize over \( j \) in the hopes that \( \min_j(V_j^*/U_j^*) \) will effectively give zero so that \( \min_j(V_j^*/U_j^*) \) will be a good estimate of \( p \). However, it seems hard to estimate \( p \) directly from a function of the ratios of \( U_j^* \)'s and \( V_j^* \)'s. The problem seems to be that the variability in \( U_j^* \) and \( V_j^* \) is so high that ratios are unstable (Gosink et al., 2007). Here, we use the same heuristic argument but apply it to stabilized forms of \( U_j^* \) and \( V_j^* \).

Specifically, we stabilize \( U_j^* \) and \( V_j^* \) by taking logarithms. Write

\[
U_j = U_j(\alpha) = \log\left(1 + \alpha U_j^*\right)
\]

\[
V_j = V_j(\alpha) = \log\left(1 + \alpha V_j^*\right)
\]

where \( \alpha > 0 \) is a scaling parameter; as \( \alpha \) increases, the degree of stabilization on the log scale decreases.

\[
R_j(\alpha) = \frac{V_j}{U_j}
\]

assumed to have finite moments \( \mu_j(\alpha) = ER_j(\alpha) \) for all \( j \) with \( K_1 \leq \mu_j \leq K_2 \), for finite \( K_1 \) and \( K_2 \), over a large range of \( \alpha \)'s (to be specified shortly).

Clearly, \( R_j \) is well defined as long as \( U_j \neq 0 \), that is, \( U_j^* \neq 0 \), meaning that no genes from the stromal sample are unexpressed. If this fails, one may ignore such genes (provided they are not too numerous) or use the symmetry of the problem to estimate \( 1 - p \). Similarly, if there are cases where \( E(R_j) = \infty \), some of the mixed tissue genes may have to be ignored to satisfy the moment conditions. In practice, these limitations have not been a problem because the difference in expression between stromal and mixed tissue samples is one of degree, not presence or absence.

For now, consider the value of \( \alpha \)

\[
\alpha^* = \arg \max_{\alpha} E \left[ \left( \frac{1}{m} \sum_{j=1}^{m} R_j(\alpha) \right)^n \right] = \arg \max_{\alpha} E(\overline{R}(\alpha))^n
\]

where the derivative on the right is with respect to \( \alpha \) and the bar on \( R \) indicates the mean over the \( m \) genes. This maximization is equivalent to minimizing the radius of curvature of the curve \( (\alpha, \overline{R}(\alpha)) \).

Given this, the true value of \( p \) can be taken as

\[
p^* = \min_j E(R_j(\alpha^*)|J = j)
\]
The reason is that \( V_j^* / U_j^* = \rho + (1 - \rho) W_j^* \). So, the smallest ratio of \( V_j^* / U_j^* \) occurs when \( W_j^* / U_j^* \) is smallest; the smallest value of \( W_j^* / U_j^* \) in real data sets is typically zero. Transforming to \( V_j / U_j \) does not change the reasoning as can be seen by expressing \( U_j^* \) and \( V_j^* \) in terms of \( U_j \) and \( V_j \) from (3) and (4). Therefore, assuming \( \rho^* \) satisfies (6) is reasonable. Now, given a good estimate of \( \alpha \), say \( \hat{\alpha} \), we can estimate \( \rho \) by

\[
\hat{\rho} = \min_j R_j(\hat{\alpha}).
\]

So, an independent and identical (IID) sample of size \( n \), \( (U_{ij}^*, V_{ij}^*) \) for \( i = 1, \ldots, n \), gives values \( (U_{ij}, V_{ij}) \) for \( i = 1, \ldots, n \) for a given \( \alpha \) and hence ratios \( R_{ij} = U_{ij} / V_{ij} \). Now,

\[
\hat{\alpha} = \arg \max_\alpha \left( \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m R_{ij}(\alpha) \right)
\]  

(7)

and

\[
\hat{\rho} = \min_j \frac{1}{n} \sum_{i=1}^n R_{ij}(\hat{\alpha})
\]  

(8)

can be used as estimators of \( \alpha^* \) and \( \rho^* \).

### 3 Consistency

Consistency of \( \hat{\alpha} \) and \( \hat{\rho} \) will be seen to follow from a uniform law of large numbers as \( n \to \infty \) (Lemma 3.1 and Proposition 3.1) and from a large deviation principle as \( m \to \infty \) (Theorem 3.1).

The form in which the uniform law of large numbers occurs is encapsulated in conditions C1 and C2. Let

\[
\bar{R}_j(\alpha) = \frac{1}{m} \sum_{j=1}^m R_{ij}(\alpha) \quad \text{and} \quad \bar{R}(\alpha) = \frac{1}{n} \sum_{i=1}^n \bar{R}_i(\alpha)
\]

so that \( \bar{R} \) is an average of IID random variables and \( \hat{\alpha} = \arg \max_\alpha \bar{R}(\alpha) \). It is seen that all \( \bar{R}_i \) are continuous as a function of \( \alpha \). Also, we can assume \( \alpha \in [\eta, 1/\eta] \) without loss, at least in practice, for some \( 1 \neq \eta > 0 \). Now consider the following.

C1: Let \( \epsilon > 0 \) and assume \( \overline{ER}(\alpha) \) exists for \( \alpha \in [\eta, 1/\eta] \), an interval large enough to contain an open set around \( \alpha^* \). Then, given a fixed value \( m \),

\[
P \left( \sup_{\alpha \in [\eta, 1/\eta]} \left| \bar{R}(\alpha) - \overline{ER}(\alpha) \right| < \epsilon \right) \to 1 \text{ as } n \to \infty.
\]

Readily verifiable conditions for C1 can be found in Andrews (1987); see also Pötscher & Prucha (1989). An excellent discussion of the strategy of proof can be found in Bierens (2005a, Chapter 6), and a particularly clear statement of sufficient conditions for C1 can be found in Theorem 3 of Bierens (2005b). Essentially, aside from technical conditions, C1 holds when expected local suprema and infima of any marginal, for example, \( \bar{R}_1 \), are bounded away from \( \infty \) and \( -\infty \), respectively.
In fact, to prove the desired consistency of $\hat{\alpha}$, we will also need a version of C1 for $R^{-1}$. Because $\alpha \in [\eta, 1/\eta]$ and the $R_i(\alpha)$’s are continuous in $\alpha$, we can write

$$[\beta_1, \beta_2] = \left[ \inf_{\alpha \in [\eta, 1/\eta]} E\left( R_1(\alpha) \right) + \xi, \inf_{\alpha \in [\eta, 1/\eta]} E\left( R_1(\alpha) \right) - \xi \right]$$

for some $\xi > 0$ small enough, where $R_1(\alpha)$ merely means $R_i(\alpha)$ for $i = 1$ because the $R_i$’s are IID. Apart from $\epsilon$, expression (9) is the range of $E(R(\cdot))$ as a function of $\alpha$. Our condition is as follows.

C2: Let $\epsilon > 0$, $\beta \in [\beta_1, \beta_2]$ where $\beta_1 > 0$ or $\beta_2 < 0$, and assume $E(R(\alpha))$ exists for $\alpha \in [\eta, 1/\eta]$. Then, given $m$,

$$P\left( \sup_{\beta \in [\beta_1, \beta_2]} |(R)^{-1}(\beta) - (E(R))^{-1}(\beta)| < \epsilon \right) \to 1 \quad \text{as} \quad n \to \infty.$$  

Informally, this means that the inverse of the mean converges to the inverse of its expectation, uniformly in $\beta$ even though $(R)^{-1}(\cdot)$ is a random function.

Because sufficient conditions for C1 are available, we give further conditions under which C1 implies C2. Taken together, the resulting conditions will enable us to state easily the other hypotheses required for consistency of $\hat{p}$ and $\hat{\alpha}$. Because $R(\alpha)$ is a random function of the form $R(\omega, \alpha)$ where $\omega \in \Omega$, the underlying measure space, we have the following.

**Lemma 3.1**
Suppose $R: \alpha \to R(\alpha)$ is a real-valued function for each $\omega \in \Omega$, is continuous in $\alpha$ almost everywhere $\omega \in \Omega$, and is uniformly continuous on compact sets, uniformly over $n$ and $m$. Suppose $E(R(\alpha)) = E(R_1(\alpha))$ has a unique maximum at $\alpha^* \in (\eta, 1/\eta)$ for all values of $m$, and suppose $\beta^* = E(R_1(\alpha^*)) \in (\beta_1, \beta_2)$ with $\beta_k = E(R_1(\alpha_k))$ for $k = 1, 2$, independently of $m$. Assume also that for all $n$ and all values of $m$, the $(R)^{-1}(\cdot)$’s are defined on $[\beta_1, \beta^*]$ and $[\beta^*, \beta_2]$. Then, for each value of $m$,

$$C1 \Rightarrow C2$$

and (10) holds even when $m$ is permitted to increase slowly.

**Proof**
Fix $m$. The probability in C2 can be written as

$$P\left( \max_{\beta \in [\beta_1, \beta^*]} \left| (R)^{-1}(\beta) - E((R)^{-1}(\beta)) \right|, \right.$$

$$\left. \sup_{\beta \in [\beta^*, \beta_2]} \left| (R)^{-1}(\beta) - E((R)^{-1}(\beta)) \right| < \epsilon \right).$$

We consider only (11) as the proof for (12) is symmetric. Because $E(R(\cdot))$ is strictly monotonic from $[\alpha_1, \alpha^*]$ onto $[\beta_1, \beta^*]$ and hence has an inverse, the first entry in the max in (11) is bounded above on the range of $E(R(\cdot))$ using...
First, consider (13). Because $\tilde{R}$ is uniformly continuous on compact sets, uniformly in $n$, $(\tilde{R})^{-1}$ is also uniformly continuous on compact sets (away from zero). So, for given $\epsilon > 0$, there exists a $\delta > 0$ such that if $|x - y| < \delta$ then $|(\tilde{R})^{-1}(x) - (\tilde{R})^{-1}(y)| < \epsilon$. Letting $x = E(\tilde{R}(\alpha))$ and $y = \tilde{R}(\alpha)$, C1 implies $|x - y| \to 0$, uniformly in $\alpha$. Now, term (13) is bounded from above by

$$P\left( \sup_{\beta \in [\beta_1, \beta^*]} \left| (\tilde{R})^{-1}(\beta) - E((\tilde{R})^{-1}(\beta)) \right| < \epsilon \right)$$

$$\leq P\left( \sup_{\alpha \in [\alpha_1, \alpha^*]} \left| (\tilde{R})^{-1}(E(\tilde{R}(\alpha))) - E((\tilde{R})^{-1}(E(\tilde{R}(\alpha)))) \right| < \epsilon \right)$$

$$+ P\left( \sup_{\alpha \in [\alpha_1, \alpha^*]} \left| (\tilde{R})^{-1}(E(\tilde{R}(\alpha))) - \alpha \right| < \epsilon \right)$$

$$+ P\left( \sup_{\alpha \in [\alpha_1, \alpha^*]} \left| (\tilde{R})^{-1}(E(\tilde{R}(\alpha))) - (\tilde{R})^{-1}(\tilde{R}(\alpha)) \right| < \epsilon \right).$$

Now, for each fixed $\delta$, C1 ensures that (17) goes to zero as $n \to \infty$. So, we can let $\delta \to 0$ slowly enough that (17) still goes to zero and the bounded convergence theorem ensures (16) also goes to zero. Terms (14) and (15) are similar, and the statement for slowly increasing $m$ is obvious: merely choose a large enough $n$ for a given $\epsilon$ and $m$.

We remark that the hypotheses of Lemma 3.1 are little more than what is minimally required for the result. However, the hypotheses are difficult to verify, specifically that the $\tilde{R}$’s are uniformly continuous on compact sets uniformly over $n$. On the other hand, the $\tilde{R}$’s are means (over $i$) of second derivatives of functions which are themselves means over genes, that is, over $j$. Because means of finitely many uniformly continuous functions are uniformly continuous, $\tilde{R}(\alpha)$ is uniformly continuous for each $n$ and hence for any finite number of $n$’s. To ensure that the uniform continuity of $\tilde{R}(\alpha)$ is uniform in $n$, it is enough to assume $\tilde{R} \to E\tilde{R}$ as $n \to \infty$ uniformly in $\alpha$ and the limit is uniformly continuous. However, this, too, is hard to verify. In practice, one looks at the curve $(\alpha, \tilde{R}(\alpha))$ for $\alpha \in [\eta, 1/\eta]$ in the unit speed parametrization and verifies it looks smooth with a unique minimum radius of curvature (Clarke et al., 2010).

Given Lemma 3.1, we can establish the consistency of $\hat{\alpha}$ for $\alpha^*$.

**Proposition 3.1**

Assume the hypotheses of Lemma 3.1 and C1 for each $m$, in particular, $\alpha^* \in (\eta, 1/\eta)$ and $\beta^* \in (\beta_1, \beta_2)$, are independent of $m$, at least for large enough $m$. Suppose that the collection of functions $\{E(\tilde{R})^{-1}(\cdot) \mid n, m \in \mathbb{N}\}$ is uniformly equicontinuous on $[\alpha_1, \alpha_2]$. Then, for any $\epsilon > 0$, we have that

$$P(|\hat{\alpha} - \alpha^*| \geq \epsilon) \to 0 \text{ as } n \to \infty.$$
Remark 3.1
Under the hypotheses of Proposition 3.1, we have that for each \( m \) (and \( n \), as we have assumed independence over subjects \( i \)), \( E\overline{R}(\alpha) \) is a monotonic function from \( [\alpha_1, \alpha^*] \) to \( [\beta_1, \beta^*] \) and is a monotonic function from \( [\alpha^*, \alpha_2] \) to \( [\beta^*, \beta_2] \). Indeed, on both domain–range pairs, \( E\overline{R}(\cdot) \) is a homeomorphism for each \( m \). Moreover, because we have independence over subjects \( i \), and having looked at data-driven examples of the behaviour of \( \overline{R}(\alpha) \), it is not unreasonable in practice to assume that \( E\overline{R}(\cdot) \) is a homeomorphism and hence invertible on the two domain–range pairs. So, it is not unreasonable to assume \( E\overline{R}(\cdot)^{-1} \) is the limit of the inverses \( \{E\overline{R}^{-1}(\cdot)\} \) on the two domain–range pairs. When this convergence is uniform, the uniform equicontinuity of \( \{E\overline{R}^{-1}(\cdot) \mid n, m \in \mathbb{N} \} \) in the statement of Proposition 3.1 holds.

Proof
Recognize that \( P(|\hat{\alpha} - \alpha^*| \geq \epsilon) \) is

\[
P\left( \left| \overline{R}^{-1} \left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \overline{R}(\alpha) \right) - \frac{E\overline{R}(\alpha)}{E} \right| \geq \epsilon \right)
\]

which can be bounded from above by the sum of two terms

\[
P\left( \left| \overline{R}^{-1} \left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \overline{R}(\alpha) \right) - \frac{E\overline{R}(\alpha)}{E} \right| \geq \epsilon/2 \right) + P\left( \left| \left( \overline{R}^{-1} \left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \overline{R}(\alpha) \right) \right) - \frac{E\overline{R}(\alpha)}{E} \right| \geq \epsilon/2 \right).
\]

This follows from the triangle inequality and the union of events bound. Because the arguments in the first term are the same, it is bounded above by

\[
P\left( \left| \overline{R}^{-1}(\beta) - \frac{E(\overline{R}(\alpha))}{E} \right| \geq \epsilon/2 \right)
\]

which goes to zero under \( C2 \).

We deal with the second term by intersecting with the event \( \left\{ \sup_{\alpha \in [\alpha_1, \alpha_2]} |\overline{R}(\alpha) - E(\overline{R}(\alpha))| \geq \epsilon \right\} \) and its complement and taking an upper bound. Now the second term is bounded from above by

\[
P\left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \left| \overline{R}(\alpha) - E(\overline{R}(\alpha)) \right| \geq \epsilon \right) \tag{19}
\]

\[
+ P\left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \left| \overline{R}(\alpha) - E(\overline{R}(\alpha)) \right| \geq \epsilon \right) \cap \left( \sup_{\alpha \in [\alpha_1, \alpha_2]} \left| \overline{R}(\alpha) - E(\overline{R}(\alpha)) \right| \geq \epsilon/2 \right) \tag{20}
\]

Term (19) goes to zero under \( C1 \).
For (20), note that for points in the underlying sample space of its event, we have that

\[ E\overline{R}(\alpha) - \epsilon \leq \overline{R}(\alpha) \leq E\overline{R}(\alpha) + \epsilon \]

and therefore

\[ \sup_{\alpha \in [\alpha_1, \alpha_2]} E\overline{R}(\alpha) - \epsilon \leq \sup_{\alpha \in [\alpha_1, \alpha_2]} \overline{R}(\alpha) \leq \sup_{\alpha \in [\alpha_1, \alpha_2]} E\overline{R}(\alpha) + \epsilon \]

which gives

\[ \left| \sup_{\alpha \in [\alpha_1, \alpha_2]} \overline{R}(\alpha) - \sup_{\alpha \in [\alpha_1, \alpha_2]} E\overline{R}(\alpha) \right| \leq \epsilon. \]

By the uniform equicontinuity of \( E(\overline{R})^{-1}(\cdot) \), we can choose \( n \) and \( m \) so large that the second event in (20) occurs with arbitrarily small probability. Hence, (18) holds. □

Given the consistency of the selection of the transformation in Proposition 3.1, we can now establish the consistency of \( \hat{p} \) for \( p^* \). Let \( \xi > 0 \) and for \( j = 1, \ldots, m \) consider the local supremum

\[ \gamma(j, \xi) = \gamma_n(j, \xi) = \sup_{\alpha \in B(\alpha^*, \xi)} \left| \frac{1}{n} \sum_{i=1}^{n} R_{ij}(\alpha) - \frac{1}{n} \sum_{i=1}^{n} \overline{R}_{ij}(\alpha^*) \right| \] (21)

where \( B(\alpha^*, \xi) \) is the interval with half-length \( \xi \) centred at \( \alpha^* \). Also, let \( M_j \) be the moment-generating function of \( -R_{1j}(\alpha^*) \) where \( \overline{R}_{ij} \) merely means \( \overline{R}_{ij} \) for \( i = 1 \) as the \( \overline{R}_{ij} \)'s are IID over \( i \). We have the following.

**Theorem 3.1**

Assume the hypotheses of Proposition 3.1 and suppose that \( \gamma(j, \xi) \to 0 \) as \( n \to \infty \), where \( \to \) means convergence in the joint probability defined in (1). Also, assume that for all \( i \) and \( j \), the \( R_{ij} \)'s are independent and that all the \( \{M_j : j = 1, \ldots, m\} \) are defined on an interval containing zero and bounded away from 1; that is, there exists \( r, c > 0 \) so that for all \( j, M_j(\cdot) \) is defined on \( (0, r) \) and \( M_j(r/2) > 1 + c \). Then, for \( p^* \in [0, 1] \), as \( n \to \infty \), there is an \( m = m(n) \to \infty \) and \( \xi = \xi(n) \to 0 \) so that

\[ \hat{p} \to p^*. \]

**Remark 3.2**

Assuming independence across the genes, that is, for different \( j \)'s, is physically unrealistic. However, the patterns of dependence are poorly understood. Consequently, it is common to pre-process gene expression data to minimize dependence across genes (Qiu et al., 2005; Klebanov & Yakovlev, 2007). Also, it is (unfortunately) typical to assume weak or minimal dependence to permit statistical analyses to proceed; see Heller & Qin (2007) for but one example. Conceptually, this also holds if the unit of interest is changed from genes to pathways.

It is also well known that both the independence and identicality assumptions can be relaxed while still maintaining a large deviation principle. This is seen in Dembo & Zeitouni (1998, Chapter 2.3), and in the large deviation principles that can be proved under various mixing conditions (e.g. Mohri & Rostamizadeh, 2010). However, the arguments are very specialized.
Proof
Let $\epsilon, \nu > 0$ and write $Y_j(\alpha) = \frac{1}{n} \sum_{i=1}^{n} R_{ij}(\alpha)$ for $j = 1, \ldots, m$. By adding and subtracting $\min_j Y_j(\alpha^*)$, intersecting with $\{|\hat{\alpha} - \alpha^*| < \nu\}$ and its complement, taking a union of events bound, and then upper bounding by dropping some conditions, we obtain an upper bound on $P(\|\hat{\alpha} - \alpha^*\| > \epsilon)$ of the form

$$P\left(\min_j Y_j(\hat{\alpha}) - \min_j Y_j(\alpha^*)\right) > \epsilon/2 \text{ and } |\hat{\alpha} - \alpha^*| < \nu$$

(22) 

$$+ P(|\hat{\alpha} - \alpha^*| \geq \nu)$$

(23) 

$$+ P\left(\min_j Y_j(\alpha^*) - \min_j EY_j(\alpha^*)\right) > \epsilon/2.$$

(24)

We begin with (22). For all $j$ and $\alpha \in B(\alpha^*, \xi)$, (21) gives

$$Y_j(\alpha^*) - \gamma(j, \xi) \leq Y_j(\hat{\alpha}) \leq Y_j(\alpha^*) + \gamma(j, \xi).$$

By standard reasoning, this leads to

$$\min_j Y_j(\alpha^*) - \max_j \gamma(j, \xi) \leq \min_j \left(Y_j(\alpha^*) - \gamma(j, \xi)\right) \leq \min_j Y_j(\hat{\alpha})$$

$$\leq \min_j \left(Y_j(\alpha^*) + \gamma(j, \xi)\right) \leq \min_j Y_j(\alpha^*) + \max_j \gamma(j, \xi).$$

Because each $\gamma(j, \xi) \to 0$ in probability, the maximum over $j = 1, \ldots, m$ also goes to zero, for each $\xi$. Therefore, (22) $\to 0$ in probability for each $m$ and $\xi$ as $n \to \infty$ as well.

Equation (23) goes to zero by Proposition 3.1.

Finally, we show that (24) $\to 0$ as $n \to \infty$. For notational convenience, redefine $Y_j$ to be evaluated at $\alpha^*$ and to be negative. That is, set $Y_j = -Y_j(\alpha^*) = -R_j(\alpha^*)$. Now (24) is $P(\|Y_j - \max_j EY_j| > \epsilon/2)$ and is bounded above by

$$P\left(\max_j Y_j - \max_j EY_j > \epsilon/2\right) + P\left(\max_j Y_j - \max_j EY_j < -\epsilon/2\right).$$

(25)

For (25), we give a version of the argument used to prove the upper bound of Theorem 1.2.6 in Deuschel & Stroock (1989); see also Stroock (1984). We see that

$$P\left(\max_j Y_j - \max_j EY_j > \epsilon/2\right) = 1 - \prod_j P\left(Y_j \leq \max_j EY_j + \epsilon/2\right) \leq 1 - \prod_j P(Y_j \leq EY_j + \epsilon/2).$$

(26)

Consider a fixed value of $m$. We show that for each $m$, there is an $N$ so that for $n > N$, $P(Y_j \leq EY_j + \epsilon/2) \to 1$ uniformly over $j = 1, \ldots, m$. To see this, observe that for each $j$, $Y_j$ is an average of $n$ independent and identically distributed random variables; write $\mu_j = EY_j$. Then for any positive $\lambda$,
because $M_j$ is the moment-generating function of any $-R_{ij}$. Therefore, any factor in (26) satisfies

$$P(Y_j \leq EY_j + \epsilon/2) \geq 1 - \min_\lambda e^{-\lambda n(\mu_j + \epsilon/2)}(M_j(\lambda))^n \geq 1 - e^{-\xi n},$$

uniformly in $j$ for some $\xi > 0$ using the hypotheses on the moment-generating functions. Hence, $\prod_j P(Y_j \leq EY_j + \epsilon/2) \to 1$ and (25) $	o 0$.

For (25), note

$$P\left(\max_j Y_j - \max_j EY_j < -\epsilon/2\right) = \prod_j P\left(Y_j < \max_j EY_j - \epsilon/2\right).$$

(27)

Consider a sequence $j_k$ such that $EY_{jk} \to \max EY_j$. Then for any $k$,

$$P\left(Y_{jk} < \max_j EY_j - \epsilon/2\right) = \lim_{k' \to \infty} P\left(Y_{jk'} < EY_{jk'} - \epsilon/2\right).$$

There also exists $K$ so that $k, k' > K$ implies $P(Y_{jk} < EY_{jk'} - \epsilon/2) \leq P(Y_{jk} < EY_{jk} - \epsilon/4)$. So there exists $\exists \xi > 0$ for which

$$P(Y_{jk} < EY_{jk} - \epsilon/4) \leq e^{-n \xi}$$

(28)

for $k > K$, by the lower-bound portion of Theorem 1.2.6 in Deuschel & Stroock (1989). (For brevity, we do not give this argument here, but see Stroock, 1984, for an accessible version.) The bound in (28) means that at least one factor in (27) can be made arbitrarily small, while the other factors can be bounded by 1. Thus, (27) $	o 0$ as $n \to \infty$.

\section{Concluding remarks}

The contribution of this paper is to identify a set of models and prove the consistency of the estimators for parameters in those models for the deconvolution of gene expression from a mixed sample. Although expressed in the context of cancer and metagenomics, our techniques should apply to any deconvolution setting using array data. To conclude, we discuss four aspects of the overall strategy.

(1) \textit{Relation of the formal hypotheses to the physical problem} Note that in our proof, given $m, \xi, \epsilon$, one lets $n$ increase first and then chooses another triple $(m, \xi, \epsilon)$, again letting $n$ increase. While this is reasonable for formal results, actual data are the reverse: it is $m$ that is large and $n$ that is small. Despite the discrepancy between mathematical hypotheses and the actual data, statistical techniques give good performance (Speed, 2010).

(2) \textit{Correcting estimates of gene expression}: Suppose the $U_{ij}$’s represent the pure stromal gene expression and the $V_{ij}$’s represent the mixed sample gene expression. Then, we can use our estimates (7) and (8) to estimate $\alpha^*$
and \( p^* \) in the context of the model (2). To detect genes that are under-expressed or over-expressed in tumour cells relative to stromal cells, let

\[
\hat{W}_{ij} = \frac{1}{1-\hat{p}} (V_{ij} - \hat{p}U_{ij})
\]  

(29)

for \( \hat{p} \in (0, 1) \). (If \( \hat{p} = 0, 1 \), then we have genes that are only expressed in stromal or tumour cells, respectively.)

Using the IID over subjects, we can let \( \mu_{\text{str}, j} = EU_{ij} \) and \( \mu_{\text{tum}, j} = EW_{ij} \) and test the \( m \) null hypotheses

\[ H_0 : \mu_{\text{str}, j} = \mu_{\text{tum}, j} \]

The “natural” test statistics are \( t_{sj} = \hat{W}_j - \bar{U}_j / SE \), where the bars indicate means over subjects and \( SE = SE(\bar{W}_j - \bar{U}_j) \) is the standard error of the numerator.

If \( p \) were known, \( t_{sj} \) would be the sample value of a paired \( t \)-test. However, because \( p \) is estimated, we would use a permutation test with a Westfall–Young correction because it, too, is based on permutations and has some optimality properties (Meinshausen et al., 2011). Rejection of \( H_0 \) would be taken as evidence of under-expression or over-expression of gene \( j \) for \( j = 1, \ldots, m \), depending on the alternative.

(3) Extension to three or more pure tissue types: Suppose the model (2) in fact were

\[
V_j^* = p_1U_j^* + p_2W_j^* + p_3Z_j^*
\]  

(30)

with \( p_1 + p_2 + p_3 = 1 \), that is, had three components. Because our method assumes in (2) that full data on \( V^* \) and \( U^* \) are available, there are two unknowns, \( W_j \) and \( p \), and our technique estimates \( p \), so we can solve for the \( W_j \)'s. To use our methods in a model such as (30), we would need to have data \( V_j^* \) and at least the sum of two components on the right, say \( U_j^* \) and \( W_j^* \), to estimate \( p_3 \). This would give \( \hat{Z}_j^* \) by an analogue of (29). Then, provided we had data on one of \( U_j^* \) and \( W_j^* \), we could use a second application of our method to find \( p_2 \) or \( p_1 \).

This has the drawback that if there are \( q \) components on the right in (30), we need \( q - 1 \) pure samples. It is not clear how to get around this problem.

(4) Interaction terms: An alternative extension would be to include an interaction term between \( U_j^* \) and \( W_j^* \) in (2). If genomic data followed Michaelis–Menten kinetics (Cornish-Bowden, 2004), this would be

\[
V_j^* = pU_j^* + p(1-p)U_j^*W_j^* + (1-p)W_j^*.
\]  

(31)

If our technique were used in this model, it would estimate \( p \) subject to controlling \( p(1-p)W_j^* + (1-p)W_j^*/U_j^* \). Any conditions that ensured \( W_j^*, W_j^*/U_j^* \to 0 \) would automatically permit out techniques here to extend to estimating \( p \) in (31). Indeed, heuristically, we have used \( W_j^*/U_j^* \to 0 \) already, so the only modifications would come from assuming \( W_j^* \to 0 \) as well. Thus, under similar conditions, our results would apply to (31) as well. The problem is that, while interaction terms are important, they are not in general of the Michaelis–Menten form. So, the utility of (31) has not been established. Research to date on the form of interactions that would be appropriate is inconclusive, but see Zhang et al. (2005) and Bishop et al. (2008) amongst others.

Acknowledgements

The authors were supported by a National Science Foundation–Defense Threat Reduction Agency grant #NR66853W (DMS-1120404).
References


Clarke, J, Seo, P & Clarke, B (2010), ‘Statistical expression deconvolution from mixed tissue samples’, Bioinformatics, 26, 1043–1049.


