Deconvolving Gene Expression Data

Max Grazier G'Sell

December 10, 2010

1 Introduction

Recent development in microarray technology have made it possible to check biological samples for the expression of many different genes (hundreds of thousands) at once. This has opened the possiblity of screening for influential genes in cases of disease succeptibility, genetic disorders, cancer, and other biological problems. In the paper by Shen-Orr et.al.[2] that inspired this project, the authors worked with blood samples from kidney transplant patients. Their goal was to determine which genes influenced the acceptance or rejection of transplant genes when they were expressed.

When using these gene expression chips, a blood sample is taken from a patient. The entire sample is lysed, and the mRNA (expressed genes) from the entire sample are processed and run on the microchip. As a result, the microchip will measure expressed genes from a mixture of several different cell types; it is difficult to physically separate the individual cell types before running them on the chip and doing so can actually alter the gene expression within the cell, obscuring the results of the experiment.

The idea behind the work of Shen-Orr et.al.[2] is that, prior to lysing the cells and measuring the gene expression, a small subsample can be taken and the distribution of cell types can be counted with a Coulter Counter, giving an estimate of the cell type distribution in the original sample. This additional information can then be used to deconvolve the gene expression data to obtain expression levels on a cell-type by cell-type basis. This can give much better resolution for seeing differences in expression in particular cell types within the sample, opening the way to interesting biological discoveries. Figure 1 shows a schematic for the data we intend to use.

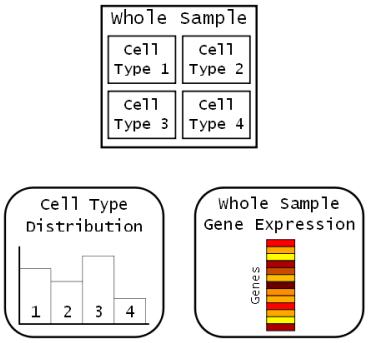

Figure 1: From the sample made up of several cell types, it is relatively easy to measure the total sample gene expression and the distribution of cell types within the sample.

1.1 Problem Setup

From the gene expression chip, we observe measurements X_{ij} of the gene expression of gene j in patient i. Suppose we have J genes and I patients. We can then write these measurements as an $I \times J$ matrix X. Each row corresponds to a particular blood sample, and each column corresponds to the gene expression of a particular gene.

Suppose that there are K dominant cell types in the mixture. Each blood sample contains a mixture of these different cell types, each with its own characteristic gene expression pattern.

Thus the expression of a particular gene j in a particular sample i is a weighted mixture of the gene expression pattern for each of the cell types present:

$$X_{ij} = \sum_{k=1}^{K} W_{ik} H_{kj}.$$

Here W_{ik} is the fraction of blood sample i made up of cell type k, so $\sum_{k=1}^{K} W_{ik} = 1$ and $W_{ik} \geq 0$. H_{kj} is the expression of gene j in cell type k.

As matrices, we can write $X = WH$, where X is $I \times J$, W is $I \times K$, and H is $K \times J$. The restrictions on W imply that it is a stochastic matrix. This has a nice geometric interpretation. The rows of X (gene expression patterns for a particular blood sample) are convex combinations of the rows of H (archetypal gene expression patterns for cell types). Therefore the rows of X lie within the convex hull of the rows of H. This will become a useful view later.

The Microarray measures all of the elements of X, and the Coulter counter can measure individual rows of the W matrix. The quantity of interest is H. If it can be estimated, then the gene expression is known on a cell-type by cell-type basis. This would allow comparison in the gene expression on the level of individual cell types between populations. Thus biological effects that depend on gene expression in a particular cell type can be observed.

In the original Shen-Orr et.al. paper[2], the corresponding rows of W were measured for every blood sample, giving the entire W matrix. This is an expensive procedure; I am interested in extending these methods to partial measurements of W.

2 Estimating H

2.1 When W is known (Shen-Orr et. al)

In the case where W is known completely (so we measure $W_{i\cdot}$ for each sample i), this problem is handled in Shen-Orr et.al.[2]. We can imagine modeling $X = WH + \varepsilon$ where ε is just uniform normal noise (for convenience). In this case we can estimate H simply by solving the multivariate regression problem $X = WH$, since both X and W are known. The closed form solution is just $H = (W^T W)^{-1} W^T X$ as we would expect. Shen-Orr. et. al.[2] carried this estimation out successfully on the kidney transplant data and were able to observe interesting differences in the H matrix.

2.2 When $\geq K$ rows of W are known

Now suppose that, due to the expense of measuring rows of W, we are able to only partially measure W. This means that we have measured the cell distribution of some of the samples, but not all of them, meaning that W has some unknown rows. In particular, let $\tilde{I}$ be the number of observed rows of W, and for now assume that $K \leq \tilde{I} < I$ (I will address the other case later).

Since $\tilde{I} \geq K$, the problem of estimating H is still identifiable. In particular, if we formed the modified matrices $\tilde{X}$ and $\tilde{W}$ containing only the rows for which W was estimated, the form of H will be exactly the same. Thus the problem becomes $\tilde{X} = \tilde{W} H$, and we can again estimate H through multiple linear regression as $H = (\tilde{W}^T \tilde{W})^{-1} \tilde{W}^T \tilde{X}$.

However, this seems to throw away useful information. Even though we have only measured W for a subset of the rows, we have measurements X for all rows. This should be useful to reduce the variance of our H estimate.

If we knew H, we could estimate the missing row of W from H and X by regression. Note that we can write $X^T = H^T W^T$. If the i^{th} row of W is missing, we can then estimate it by linear regression: $W_{i\cdot}^T = (HH^T)^{-1} H X_{i\cdot}^T$, or equivalently $W_{i\cdot} = X_{i\cdot} H^T (HH^T)^{-1}$.

With this in mind, I propose the following EM algorithm for estimating H that takes into account the additional known rows of X:

(0) Initialize $H^{(0)} = (\tilde{W}^T \tilde{W})^{-1} \tilde{W}^T \tilde{X}$, the estimate using only known rows of W.

(ℓ) Repeat until convergence:

(a) Estimate the missing rows i of W, $W_{i\cdot}^{(\ell)} = X_{i\cdot} (H^{(\ell-1)})^T (H^{(\ell-1)} (H^{(\ell-1)})^T)^{-1}$. Form

$W^{(\ell)}$.

(b) Estimate $H^{(\ell)}$ by $H^{(\ell)} = ((W^{(\ell)})^T W^{(\ell)})^{-1}(W^{(\ell)})^T X$.

This should converge to values of H and W that are more consistent with the overall X matrix.

2.3 When $< K$ rows of W are known

It is also interesting to consider the same setup, but now with $\tilde{I} < k$. In this case, the problem of finding H is unidentifiable. Thus the EM algorithm above will fail. The data simply do not contain enough information to estimate H without further restrictions on its form.

For completely unknown W, Rob Tibshirani suggested using Archetypal Analysis [1] to attempt to find both W and H. Archetypal Analysis imposes the additional assumption/restriction that $H = BX$, where B is again a stochastic matrix. This constrains the problem enough that it is possible to solve.

Geometrically, the additional criterion that $H = BX$ is similar to our original statement $X = WH$. $X = WH$ says that the rows of X are in the convex hull of the rows of H. The constraint $H = BX$ attempts to put the rows of H in the convex hull of X. These cannot both be satisfied unless the convex hulls are identical. However, we are fitting this model assuming errors, so the combined criteria will choose H so the convex hulls of the rows of H and X are close. This should cause the rows of H to represent extreme points of X, without deviating too far in the convex hulls they represent.

Whether this constraint is appropriate is unclear, but it seems somewhat reasonable. It makes more sense if a reduced weight is put on the $H = BX$ constraint, allowing greater deviations of the H rows from the convex hull of X. This would let the convex hull of H go outside the convex hull of X, which is a bit more what we would expect. Professor Tibshirani found that this seemed to give more stable results on the true data.

There is one major problem with decomposing $X = WH$ with W unknown: the solution is invariant to permutations of the K rows of H and columns of W. Practically, it is impossible to figure out which of the rows of H correspond to each of the cell types of interest. Thus, to say anything meaningful, it seems that we need at least some measurements of W to make these archetypal points unambiguous.

I propose a combination of the archetypal analysis approach with the known information on W. Thus seek W, H, B to minimize the criterion

$$L(H, W, B) = ||X - WH||_2^2 + \gamma||H - BX||_2^2,$$

with the additional constraint that W is consistent with $\tilde{W}$ for the known rows (could be captured by additional $\gamma_2||\tilde{W} - W_{\tilde{I}}||_2^2$ penalty if we allow error in $\tilde{W}$). We expect that we should choose $\gamma < 1$. This problem is biconvex and so should be solvable by iterative convex optimization.

I will not include the results for $\tilde{I} < k$ in this paper for two reasons. One is the lack of space in this writeup. The other is that the problem is less interesting. In typical problems, k is quite small ($k = 5$ in the Shen-Orr paper). Since it is necessary to measure some of the rows of W to disambiguate the cell types, there is little reason not to measure at least k of them, making this method less interesting

3 Results

I tested the first algorithm on two different data sets. The first was simulated data, to see how the algorithm performed in an ideal setting. The second was the same kidney data used in the Shen-Orr paper[2].

3.1 Simulation Results

For the simulated data, I used a small, simple setup, so the results could be visualized. The simulated data were generated as follows. We have $I = 20$ samples, $J = 50$ genes, and $K = 3$ cell types. The true W matrix of weights is

generated uniformly by stickbreaking for each row, and the true archetypes H are generated $U[0, 10]$ for each entry in H. Then the actual observed data X are computed as $WH+\varepsilon$ where $\varepsilon \sim N(0, 1)$.

For different values of $\tilde{I}$, $3 \leq \tilde{I} \leq 20$, we use the full X matrix and $\tilde{I}$ rows of the W matrix to estimate H by the above algorithm. We can look at this in several ways. First, to get an idea how well our solution for H approximates the true H, we can compute the average Frobenius norm between the true and the approximate H, over the range of possible $\tilde{I}$. We obtain the plot shown in Figure 2. For very low values of $\tilde{I}$, we

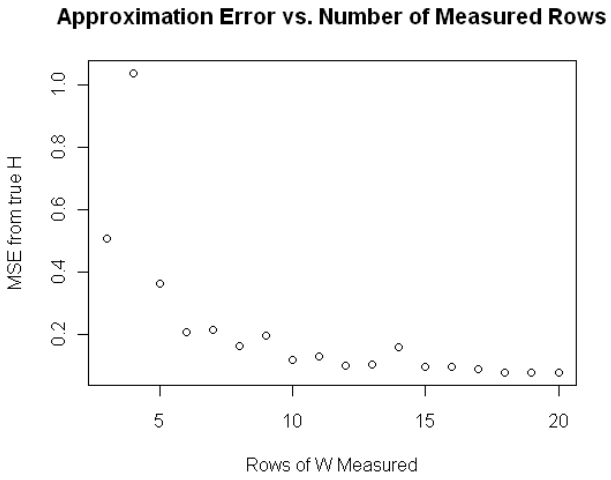

Figure 2: Distance of approximate H from true H versus number of rows of W measured.

see that H is not recovered well. However, once $\tilde{I}$ is about 6 or so, the matrix H is recovered nearly as well as in the case where W is fully measured ($\tilde{I} = 20$, the last point on the graph).

We can try to get a graphical sense of how well this worked, since the simulation used a small enough number of genes that we can visualize them explicitly. In Figure 3, we see a visual representation of gene expression by cell type. The left plot is the true (unknown) H, and the right plot is the estimated H recovered with $\tilde{I} = 12$. We see that recovery is not perfect, but there is a very strong correspondance between expression

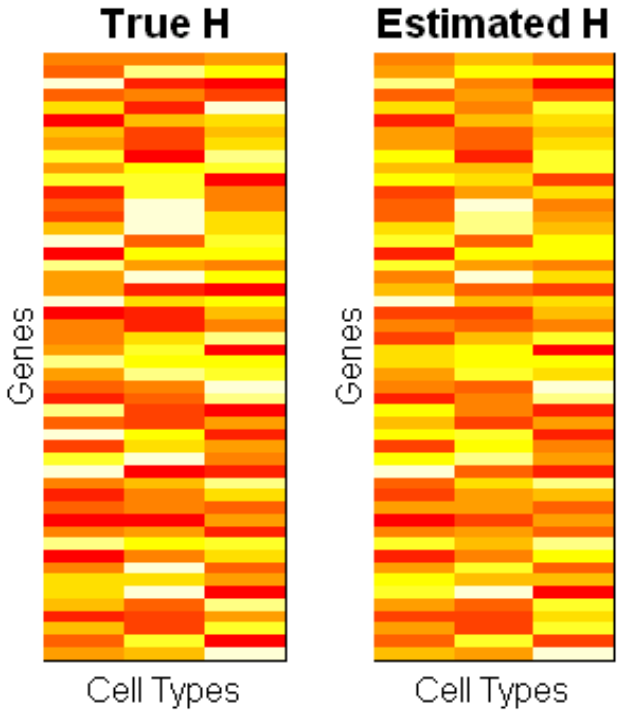

Figure 3: Expression by cell type (H) for the true H and the H recovered with $\tilde{I} = 12$, corresponding to measurement of half of the rows of W. (Using simulated data.)

levels in the two matrices. This suggests that it may indeed be possible to recover H from only partial measurements on W.

3.2 Kidney Data Results

We next try this method on the true blood sample data used in the Shen-Orr paper. We will look only at one of the study groups, In this data set, we have a sample size of $I = 15$, $K = 5$ cell types, and $J = 54675$ genes. We clearly cannot represent this number of genes graphically, but we can look at the average Frobenius norm for varied numbers of rows of W, as in Figure 2. This plot is shown in Figure 4. Since this is real instead of simulated data, we do not have a ground truth for H. Instead, we use the H estimated with all of W as the ground truth (corresponding to H the way it was estimated in the Shen-Orr paper), and see how well we approxi-

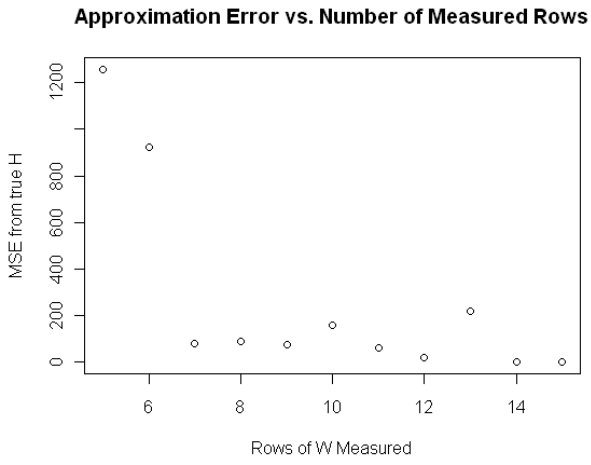

Figure 4: Distance of approximate H from the H estimated with all the rows, versus number of rows of W measured.

mate this H when measuring fewer rows of W.

Looking at the plot, the results are not quite as dramatic as in the simulated case. Nevertheless, we see that when we measure at least half of the rows of W or so, we do a reasonable job of reconstructing H.

4 Conclusion

We have proposed a method for estimating the gene expression in each cell type (H), requiring measurement of the cell type distribution (W) for only a subset of the samples. In both simulation and real data, we see that this does a reasonable job of approximating H while requiring measurement of significantly fewer rows of W. This is encouraging, opening the possibility for cheaper and faster deconvolution of data of this sort.

This could be continued in several directions. It would be interesting to try using these methods to compare population expression, as in the Shen-Orr paper, to try to identify cell types that are responsible for differences in biological and medical behavior. This could be a better metric than L^2 distance for seeing whether the method is actually viable. There is also room for work in the low $\tilde{I}$ area, as suggested by the proposed biconvex estimation problem above.

5 Acknowledgements

Thank you to Rob Tibshirani for suggesting and discussing the problem, and for providing me with the data from the Shen-Orr et.al. paper.

6 References

[1] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):pp. 338–347, 1994.

[2] Shai S. Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, Mark M. Davis, and Atul J. Butte. Cell type-specific gene expression differences in complex tissues. *Nature methods*, 7(4):287–289, April 2010.