Completion of missing entries in matrices and tensors

Shmuel Friedland
Univ. Illinois at Chicago

Department of Statistics, The University of Chicago,
September 12, 2011

Overview

- Introduction
- DNA Micorarrays
- Recovery methods of missing entries in DNA Micorarrays

Introduction

In many instances in measuring multidimensional data, as matrices and tensors, one confronts the following problems: noisy data, missing entries and data reduction.
There are many statistical and mathematical methods to deal with these problems. In this talk we survey some of the known methods and expand on the methods that the speaker was working on.

DNA Microarrays: I

A DNA microarray (also commonly known as gene chip, DNA chip, or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10-12 moles) of a specific DNA sequence, known as probes (or reporters). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation.

DNA Microarrays: II

In standard microarrays, the probes are synthesized and then attached via surface engineering to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are colloquially known as an Affy chip when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. Alternatively, microarrays can be constructed by the direct synthesis of oligonucleotide probes on solid surfaces. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system.

DNA Microarrays: III

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes (see uses and types section). Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost (see fabrication section). Additional factors for microarray experiments are the experimental design and the methods of analyzing the data (see Bioinformatics section).

DNA Microarrays: IV

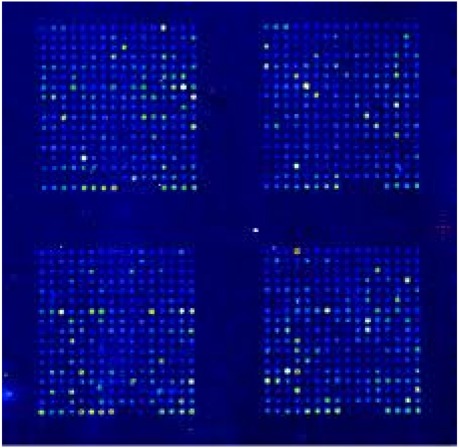


Figure: Microarays raw data

DNA Microarrays: V

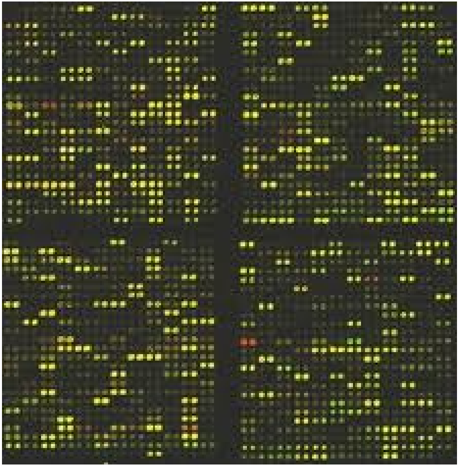


Figure: Microarays processed data

Missing entries in DNA Microarrays

During the laboratory process, some spots on the array may be missing due to various factors (for example, machine error.) Because it is often very costly or time consuming to repeat the experiment, molecular biologists, statisticians, and computer scientists have made attempts to recover the missing gene expressions by some ad-hoc and systematic methods.

Gene expression matrix

$$E = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ g_{j1} & g_{j2} & \cdots & g_{jm} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1^\top \\ \mathbf{g}_2^\top \\ \vdots \\ \mathbf{g}_j^\top \\ \vdots \\ \mathbf{g}_n^\top \end{bmatrix} = [\mathbf{c}_1\ \mathbf{c}_2\ \ldots\ \mathbf{c}_m] \in \mathbb{R}^{n\times m}$$

$$\mathbf{g}_j^\top := (g_{j1}, g_{j2}, ..., g_{jm}),\ j = 1, ..., n, \mathbf{c}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ji} \\ \vdots \\ g_{ni} \end{bmatrix},\ i = 1, ..., m.$$

$\mathbf{g}_j^\top$ relative expression levels of j^{th} gene in m experiments.

$\mathbf{c}_i$ relative expression levels of n genes in i^{th} experiment

$n \gg m$

Missing entries problem in DNA Microarrays

$\mathcal{N} \subset [n] := \{1, \ldots, n\}$ the set of rows of E that contain at least one missing entry.

For each $j \in \mathcal{N}^c := [n] \backslash \mathcal{N}$, the gene $\mathbf{g}_j^\top$ has all of its entries.

n' denote the size of $\mathcal{N}^c$, i.e. the size of $\mathcal{N}$ is $n - n'$.

Problem: complete the missing entries of each $\mathbf{g}_j^\top$, $j \in \mathcal{N}$,

under some assumptions.

Common methods of recovery

- Zero replacement method;
- Row sum mean;
- Baesian Principal Component Analysis; [1]
- Clustering analysis methods such as K-nearest neighbor clustering, hierarchical clustering [2], KNNimpute, - [2].
- FRAA and IFRAA; [4, 3]
- Least square imputation methods; [1];
- Local least squares imputation method (LLS) [2];
- Projection onto convex sets methods (POCS) [1]
- SVDimpute - Singular Value Decomposition (which is closely related to Principal Component Analysis) [2]

Short descriptions of KNNimpute and LLS

KNNimpute and LLS are local methods, which use similarity structure of the data to impute the missing values

KNNimpute uses the weighted averages of the K-nearest uncorrupted neighbors.

LLS has two versions to find similar genes whose expressions are not corrupted: the L_2-norm and the Pearson's correlation coefficients. After a group of similar genes C are identified, the missing values of the gene are obtained using least squares applied to the group C.

In these two methods, the recovery of missing data is done independently, i.e. the estimation of each missing entry does not influence the estimation of the other missing entries.

Short description of BPCA

BPCA is a global method consisting of three components. First, principal component regression, which is basically a low rank approximation of the data set is performed. Second, Bayesian estimation, which assumes that the residual error and the projection of each gene on principal components behave as normal independent random variables with unknown parameters, is carried out. Third, Bayesian estimation follows by iterations based on the expectation-maximization (EM) of the unknown Bayesian parameters.

Short descriptions of SVDimpute, FRAA and IFRAA

These methods are intimately related to Singular Value Decomposition

SVD

Singular Value Decomposition - SVD

$A = U\Sigma V^\top \in \mathbb{R}^{n\times m}$

$$\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_{\min(m,n)}) := \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \sigma_n \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \in \mathbb{R}^{n\times m}$$

$\sigma_1 \ge \ldots \ge \sigma_r > 0 = \sigma_i, i > r = \mathrm{rank}\, A$

$U = [\mathbf{u}_1 \ldots \mathbf{u}_n] \in \mathbb{O}(n), \quad V = [\mathbf{v}_1 \ldots \mathbf{v}_m] \in \mathbb{O}(m)$

$a^\dagger = a^{-1}$ if $a \neq 0$, $a^\dagger = 0$ if $a = 0$

$A^\dagger := V\,\mathrm{diag}(\sigma_1^\dagger,\ldots,\sigma_{\min(m,n)}^\dagger)U^\top$

Best rank k-approximation

For $k \le r = \mathrm{rank}\, A$: $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k \times k}$,
$U_k = [\mathbf{u}_1 \ldots \mathbf{u}_k] \in \mathbb{R}^{m \times k}$, $V_k = [\mathbf{v}_1 \ldots \mathbf{v}_k] \in \mathbb{R}^{n \times k}$
$A_k := U_k \Sigma_k V_k^\top$ is the best rank k approximation in Frobenius and operator norm of A

$$\min_{B \in \mathcal{R}(m,n,k)} ||A - B||_F = ||A - A_k||_F \quad (\|A\|_F^2 = \mathrm{tr}(AA^\top)).$$

Reduced SVD $A = U_r \Sigma_r V_r^\top$ where $(r \ge)\ \nu$ numerical rank of A if

$$\frac{\sum_{i \ge \nu+1} \sigma_i^2}{\sum_{i \ge 1} \sigma_i^2} \approx 0, (0.01).$$

A_ν is a noise reduction of A. Noise reduction has many applications in image processing, DNA-Microarrays analysis, data compression.
Full SVD: $O(mn \min(m, n))$, k- SVD: $O(kmn)$.

Minimal characterization of sum of squares of singular values

$\sigma_1^2(A) \geq \sigma_2(A)^2 \geq \ldots$ are the eigenvalues of $AA^\top$ and $A^\top A$.

Ky-Fan characterization

$$\sum_{i=\nu+1}^{m} \sigma_i(A)^2 = \min_{[\mathbf{x}_{\nu+1} \ldots \mathbf{x}_m] \in \mathbb{O}(m, m-\nu)} \sum_{i=\nu+1}^{m} (A\mathbf{x}_i)^\top (A\mathbf{x}_i)$$

$\mathbb{O}(m, k) \subset \mathbb{R}^{m \times k}$ all matrices with k orthonormal columns

SVDimpute [2, 2]

First, replace the missing values with 0 or with values computed from another method. Call the estimated matrix E_p, where $p = 0$.

Find the l_p significant singular values of E_p, and let E_{p,l_p} be the filtered part of E_p. Replace the missing values in E by the corresponding values in E_{p,l_p} to obtain the matrix E_{p+1}.

Continue this process until E_p converges to a fixed matrix (within a given precision). This algorithm takes into account implicitly the influence of the estimation of one entry on the other ones. But it is not clear if the algorithm converges, nor what are the features of any fixed point(s) of this algorithm.

Fixed Rank Approximation Algorithm (FRAA): I

$\Omega \subset \{1,\ldots,n\} \times \{1,\ldots,m\}$ missing entries set.

Set $g_{ij} = 0$ if $(i,j) \in \Omega$ to obtain $E \in \mathbb{R}^{n\times m}$.

$\mathcal{X}$ are all $X = [x_{ij}] \in \mathbb{R}^{n\times m}$ where $x_{ij} = 0$ if $(i,j) \notin \Omega$.

Assume that the completed matrix of the experiment should have the numerical rank ν. Then we complete the entries by solving the problem:

$$(1) \quad \min_{X\in\mathcal{X}} \sum_{i=\nu+1}^{m} \sigma_i^2(E+X) = \min_{X\in\mathcal{X}} \sum_{i=\nu+1}^{m} \lambda_i((E+X)^\top(E+X))$$

FRAA: II

Fixed Rank Approximation Algorithm: [4]

$G_p \in \mathcal{X}$ is p^{th} approximation to a solution of optimization problem (1).

Let $B_p := (E + G_p)^\top (E + G_p)$

Find an orthonormal set of eigenvectors for B_p, $\mathbf{v}_{p,1}, \ldots, \mathbf{v}_{p,m}$.

Then G_{p+1} is a solution to the following minimum
of a convex nonnegative quadratic function

$$\min_{X \in \mathcal{X}} \sum_{q=l+1}^{m} ((E + X)\mathbf{v}_{p,q})^\top ((E + X)\mathbf{v}_{p,q})$$

FRAA: III

Flow chart of the algorithm:

Fixed Rank Approximation Algorithm (FRAA)
Input: integers $m, n, L, iter$, the locations of non-missing entries $\mathcal{S}$, initial approximation G_0 of $n \times m$ matrix G.
Output: an approximation G_{iter} of G.
for $p = 0$ **to** $iter - 1$
- Compute $B_p := (E + G_p)^\top (E + G_p)$ and find an orthonormal set of eigenvectors for B_p, $\mathbf{v}_{p,1}, \ldots, \mathbf{v}_{p,m}$.
- G_{p+1} is a solution to the minimum problem (1) with
$\nu = L - 1 = l$.

FRAA: IV

Let $f_l(X) := \sum_{i=\nu+1}^{n} \sigma_i^2(A+X)$.

Then $f_l(G_p) \geq f_l(G_{p+1})$. $G_p, p = 1, \ldots$ converges to a critical point $\tilde{G}$.

FRAA gives a good approximation of $\tilde{G}$. In many simulations $\tilde{G} = G^*$.

FRAA is an adaptation of an algo for IEP:

Inverse Eigenvalue Problem:
Find the values of the missing entries of G such that the nonnegative definite matrix $G^\top G$ will have $m - l$ smallest eigenvalues equal to zero.
IEP appear often in engineering. See [5]

FRAA is a robust algorithm which performs good, but not as well as KNNimpute, BPCA and LSSimpute.
All other algo reconstruct the missing values of each gene from similar genes.

Fixed Rank Approximation Algorithm (IFRAA)

First use FRAA to find a completion G.

Then use a cluster algorithm

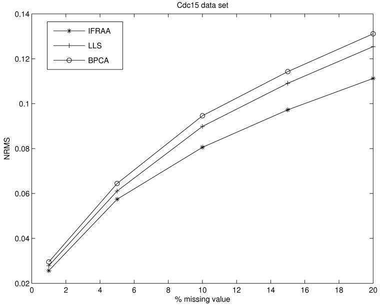
(We used K-means repeating & refining cluster size),
to find a reasonable number of clusters of similar genes,

each cluster is a relatively smaller matrix having an effective low rank.

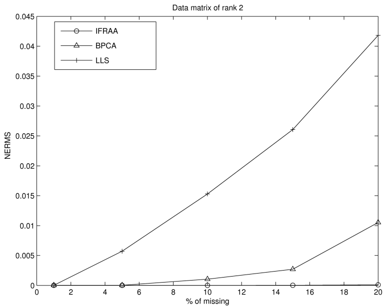
For each cluster of genes apply FRAA separately to recover the missing entries in this cluster [3].

These results suggest that IFRAA has a potential for being an effective algorithm to recover blurred spots in digital images.

SIMULATIONS 1



SIMULATIONS 2



The performance of the BCPA, IFRAA and LLS

algorithms depends on the unknown distribution of missing position of the entries.

Table: Comparison of NRMSE for three methods: IFRAA, LLS and BPCA for actual missing values distribution for three gene expression data sets with different percentage of missing values.

Data sets	IFRAA	LLS	BPCA
Cdc15 data set %0.81 missing	0.0175	0.0200	0.0216
Evolution data set %9.16	0.0703	0.0969	0.1247
Calcineurin data set %3.68	0.0421	0.0445	0.0453

Missing entries for 3-tensors

$\mathcal{T} = [t_{i,j,k}]_{i=j=k=1}^{n,m,l} \in \mathbb{R}^{n\times m\times l}$.

$\Omega \subset \{1,\ldots,n\} \times \{1,\ldots,m\} \times \{1,\ldots,l\}$ missing entries set

Simple solution: Assume $1,\ldots,n$ are genes

Unfold $\mathcal{T}$ in direction 1 to get the matrix $E = [g_{i(j,k)}] \in \mathbb{R}^{n\times(ml)}$ where $g_{i(j,k)} = t_{i,j,k}$.

Apply your favorite completion algorithm for matrices

(p, q, r)-approximation of 3-tensors

$\mathbf{U} \subset \mathbb{R}^n, \mathbf{V} \subset \mathbb{R}^m, \mathbf{W} \subset \mathbb{R}^l$ of dimensions p, q, r respectively

with orthonormal bases $[\mathbf{u}_1, \ldots, \mathbf{u}_p], [\mathbf{v}_1, \ldots, \mathbf{v}_q], [\mathbf{w}_1, \ldots, \mathbf{w}_r]$

$P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T}) = \sum_{i=j=k}^{p,q,r} \langle \mathcal{T}, \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k \rangle \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k$

$(\langle \mathcal{T}, \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \rangle = \sum_{i=j=k=1}^{n,m,l} t_{i,j,k} x_i y_j z_k)$

$\|\mathcal{T}\|_{HS}^2 := \|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})|_{HS}^2 + \|P_{(\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W})^\perp}(\mathcal{T})|_{HS}^2$

$(\|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})|_{HS}^2 := \sum_{i=j=k=1}^{p,q,r} \langle \mathcal{T}, \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k \rangle^2)$

(Best) (p, q, r)-approximation $P_{\mathbf{U}^\star\otimes\mathbf{V}^\star\otimes\mathbf{W}^\star}(\mathcal{T})$:

$\arg\max \|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS} = \arg\min \|P_{(\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W})^\perp}(\mathcal{T})\|_{HS}$

Methods to find (p, q, r)-approximation

Higher Order Singular Value Decomposition HOSVD

Unfold in direction 1 and find p-truncated SVD approximation.
$\mathbf{U}$-the subspace spanned by first p-left singular vectors.
Do similarly for $\mathbf{V}, \mathbf{W}$.

Alternating Least Squares Method:

Fix $\mathbf{V}_0, \mathbf{W}_0$, e.g. use HOSVD.
Find $U_1 := \arg\max\{\mathbf{U}, \|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS}$.
(Equivalent to finding of the first p-eigenvectors of corresponding nonnegative definite matrix.)

Fix $\mathbf{U}_1, \mathbf{W}_0$ and find $\mathbf{V}_1$, then fix $\mathbf{U}_1, \mathbf{V}_1$ and find $\mathbf{W}_1$
Continue the algorithm
In each step of the algorithm the value of $\|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS}$ increases
Convergence to a critical point, which is a semi-local maximum

Fixed Rank Approximation Algorithm for Tensors

$\Phi_\Omega \subset \mathbb{R}^{n\times m\times l}$ all tensors $\mathcal{X} = [x_{i,j,k}] \in \mathbb{R}^{n\times m\times l}$ with $x_{i,j,k} = 0$ if $(i,j,k) \notin \Omega$.

$\mathcal{T} = [t_{i,j,k}] \in \mathbb{R}^{n\times m\times l}$, $t_{i,j,k} = 0$ if $(i,j,k) \in \Omega$.

$\mathcal{X}_0$ an approximation of completed errors

Assume $\mathcal{X}_s$ given.

Find (p,q,r)-approximation of $\mathcal{T} + \mathcal{X}_s$ with corresponding subspaces $\mathbf{U}_s, \mathbf{V}_s, \mathbf{W}_s$.

Then $\mathcal{X}_{s+1} := \arg\min\{\|P_{(\mathbf{U}_s\otimes\mathbf{V}_s\otimes\mathbf{W}_s)^\perp}(\mathcal{T}+\mathcal{X})\|_{HS}, \mathcal{X}, \in \Phi\}$.

$\mathcal{X}_s$ converges to a critical semi-local maximum

References 1

- T.H. Bø, B. Dysvik and I. Jonassen, LSimpute: accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research*, 32 (2004), e34
- H. Chipman, T.J. Hastie and R. Tibshirani, Clustering micrarray data in: T. Speed, (Ed.), Statistical Analysis of Gene Expression Microarray Data, , Chapman & Hall/CRC, 2003 pp. 159-200.
- S. Friedland, M. Kaveh, A. Niknejad, H. Zare, *An algorithm for missing value estimation for DNA microarray data*, Proc. ICASSP, 2006.
- S. Friedland, A. Niknejad and L. Chihara, A Simultaneous Reconstruction of Missing Data in DNA Microarrays, *Linear Algebra Appl.*, 416 (2006), 8-28.
- S. Friedland, J. Nocedal and M. Overton, The formulation and analysis of numerical methods for inverse eigenvalue problems, SIAM J. Numer. Anal. 24 (1987), 634-667.

References 2

- X. Gan, A.W.-C. Liew and H. Yan, Missing Microaaray Data Estimation Based on Projection onto Convex Sets Method, *Proc. 17th International Conference on Pattern Recognition*, 2004
- H. Kim, G.H. Golub and H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, *Bioinformatics* 21 (2005), 187-198.
- Amir Niknejad, *Application of Singular Value Decompositions to DNA Microarrays*, Ph.D. thesis, UIC, 2005, http://www2.math.uic.edu/∼friedlan/thesis9.19.05.pdf
- A. Niknejad and S. Friedland, *APPLICATIONS OF LINEAR ALGEBRA TO DNA MICROARRAYS*, VDM Verlag Dr Müller Aktiengesellschaft&Co.KG, Germany, 2009, ISBN: 978-3-639-17994-1

References 3

S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii, A Baesian missing value estimation method for gene expression profile data, *Bioinformatics* 19 (2003), 2088-2096

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, Missing value estimation for DNA microarrays, Bioinformatics 17 (2001), 520-525.