Plaid Models and Microarrays

Art Owen

owen@stat.stanford.edu

Stanford University

Joint work with:

Laura Lazzeroni

Stanford University

laura@osiris.stanford.edu

Expression data

For Genes $i = 1, \ldots, n$

And Samples $j = 1, \ldots, p$

Y_{ij} measures expression level

Samples

From different organs/individuals/times, etc.

Genes

Many or all of the organism's genes

Expression

Activity level of gene i in sample j

Starting points

1. Eisen, Spellman, Brown, Botstein: PNAS (1998)
2. Hastie, Tibshirani, Eisen, Brown, Ross, Scherf, Weinstein, Alizadeh, Staudt, Botstein: S.U. Tech. Report (2000)

Transposable Data

Observations	Variables	Data	Dimension
Genes	Samples	Y	$n \times p$
Samples	Genes	Y^T	$p \times n$
Movies	Viewers	Y	$n \times p$
Viewers	Movies	Y^T	$p \times n$
Words	Documents	Y	$n \times p$
Documents	Words	Y^T	$p \times n$

Good statistics problems:

1. (When) should model be symmetric?
2. Can't have both $n/p \to \infty$ and $p/n \to \infty$
3. How best to bootstrap?
4. How to use row/col specific covariates?

Managing the data

n is usually large

p can be large

. . . graphical methods

Eisen et. al. (1998) data

2467 yeast genes

79 samples

multiple experiments

Modeling Approach

Try:

$$
\begin{aligned}
Y_{ij} &\doteq \mu_0 + \sum_{k=1}^{K} \rho_{ik}\, \kappa_{jk}\, \mu_k \\
\rho_{ik} &\in \{0, 1\} \\
\kappa_{jk} &\in \{0, 1\}
\end{aligned}
$$

Interpretation:

μ_0 is a background level

There are K "layers", with levels μ_k

ρ_{ik} for gene membership

κ_{jk} for sample membership

$\mu_k > 0$, for upregulation

$\mu_k < 0$, for downregulation

Bigger models

$$Y_{ij} \doteq \mu_0 + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\mu_k$$

$$Y_{ij} \doteq \mu_0 + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\left[\mu_k + \alpha_{ik}\right]$$

$$Y_{ij} \doteq \mu_0 + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\left[\mu_k + \beta_{jk}\right]$$

$$Y_{ij} \doteq \mu_0 + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\left[\mu_k + \alpha_{ik} + \beta_{jk}\right]$$

Subject to

$$\sum_{i=1}^{n} \rho_{ik}\alpha_{ik} = 0, \quad \forall k$$

$$\sum_{j=1}^{p} \kappa_{jk}\beta_{jk} = 0, \quad \forall k$$

Anova-lets, but without the orthogonality

Geometry of a layer

Include β_{jk} but not α_{ik}

Drop subscript k

Let

$$m = \sum_{i=1}^{n} \rho_i$$

$$q = \sum_{j=1}^{p} \kappa_j$$

$$C = \mu + \beta_j \in \mathbb{R}^q$$

Like a cluster of m genes around $C \in \mathbb{R}^q$

$m \leq n$ some, maybe not all, genes

$q \leq p$ some, maybe not all, samples

$\mu + \beta_j$ gives an "expression pattern"

Importance of sample j given by $|\mu + \beta_j|$

Adding layers: lets genes be in multiple clusters

Converse: get cluster of samples wrt some genes

More geometry

Consider $\mu + \alpha_i + \beta_j$

Genes i cluster around a line through

$$C_G \quad = \quad \mu + \beta_j \in \mathbb{R}^{\mathrm{q}}$$

Samples j cluster around a line through

$$C_S \quad = \quad \mu + \alpha_i \in \mathbb{R}^{\mathrm{m}}$$

Most important/typical genes: large $\left|\mu + \alpha_i\right|$

Even Bigger models

1. Write

$$\mu_0 + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\Big[\mu_k + \alpha_{ik} + \beta_{jk} + \lambda_k\rho_{ik}\kappa_{jk}\Big]$$

2. Incorporates Tukey's 1 df for non-additivity
3. Clusters genes around a more general line in $\mathbb{R}^q$
4. We can mix/match layer types.
5. We can replace the background μ_0 by a model layer.

SVD and others

$$Y_{ij} \doteq \sum_{k=1}^{K} \sigma_k u_{ik} v_{jk}$$

Method	σ_k	u_{ik}	v_{ik}	Also:
SVD	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$u_{\cdot k}^T u_{\cdot k'} = \delta_{kk'}$
SDD	$\mathbb{R}$	$\{0, \pm 1\}$	$\{0, \pm 1\}$	
NND	1	$[0, \infty)$	$[0, \infty)$	
VQ	1	$\{0, 1\}$	$\mathbb{R}$	$\sum_k u_{ik} = 1$
VQ	1	$\mathbb{R}$	$\{0, 1\}$	$\sum_k v_{jk} = 1$
Shave	1	$\{0, \pm 1\}$	$\mathbb{R}$	
ADDCL	$\mathbb{R}$	$\{0, 1\}$	$\{0, 1\}$	$n = p$
Plaid	$\star$	$\{0, 1\}$	$\{0, 1\}$	

*Plaid replaces σ_k by a model

Algorithm

Seek small value of

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij}-\sum_{k=0}^{K}\rho_{ik}\kappa_{jk}\theta_{ijk}\right)^2$$

Where

$$\begin{aligned}\rho_{ik},\kappa_{jk} &\in \{0,1\}, \quad \text{and,}\\ \theta_{ijk} &= \mu_k\\ &or \quad \mu_k+\alpha_{ik}\\ &or \quad \mu_k+\beta_{jk}\\ &or \quad \mu_k+\alpha_{ik}+\beta_{jk}\end{aligned}$$

1. Likely to be NP-hard . . . even clustering is
2. We pick one layer at a time . . .
3. . . . using an interior point algorithm
4. Larger clusters are more attractive
5. Clusters near background not attractive

Finding one layer

Residual:

$$Z_{ij} = Y_{ij} - \sum_{r=0}^{k-1} \rho_{ir}\kappa_{jr}\theta_{ijr}$$

Drop k and write:

$$Q = \frac{1}{2}\sum_i \sum_j \Big(Z_{ij} - \rho_i\kappa_j\theta_{ij}\Big)^2$$

We want to min Q over θ, ρ, κ

1. Start with arbitrary $\rho_i, \kappa_j \in (0, 1)$
2. Update θ_{ij} given ρ_i and κ_j
3. Update ρ_i and κ_j given θ_{ij}

Alternate 2, 3 above, but:

1. Keep ρ_i, κ_j away from 0 and 1 early on
2. Force ρ_i, κ_j to 0 or 1 later

Fuzzy anova

Minimize:

$$\frac{1}{2}\sum_i\sum_j\left(Z_{ij}-\rho_i\kappa_j\left[\mu+\alpha_i+\beta_j\right]\right)^2$$

Subject to:

$$0 \quad = \quad \sum_i \rho_i^2\alpha_i \quad = \quad \sum_j \kappa_j^2\beta_i$$

By taking:

$$\begin{aligned}
\mu &= \frac{\sum_i\sum_j \rho_i\kappa_j Z_{ij}}{\left(\sum_i \rho_i^2\right)\left(\sum_j \kappa_j^2\right)} \\
\alpha_i &= \frac{\sum_j (Z_{ij}-\mu\rho_i\kappa_j)\kappa_j}{\rho_i\sum_j \kappa_j^2} \\
\beta_j &= \frac{\sum_i (Z_{ij}-\mu\rho_i\kappa_j)\rho_i}{\kappa_j\sum_i \rho_i^2}
\end{aligned}$$

Updating ρ_i and κ_j

Minimize:

$$\frac{1}{2}\sum_i\sum_j\left(Z_{ij}-\rho_i\kappa_j\left[\mu+\alpha_i+\beta_j\right]\right)^2$$

Let:

$$\begin{aligned}\theta_{ij} &= \mu+\alpha_i+\beta_j\\ \rho_i &= \frac{\sum_j\theta_{ij}\kappa_j Z_{ij}}{\sum_j\theta_{ij}^2\kappa_j^2}\\ \kappa_j &= \frac{\sum_i\theta_{ij}\rho_j Z_{ij}}{\sum_i\theta_{ij}^2\rho_i^2}\end{aligned}$$

Notes

1. The α_i update only uses gene i's data
2. The ρ_i update only uses gene i's data
3. Avoids $O(n^2)$ costs
4. Similarly for β_j, κ_j and $O(p^2)$

Some details

Starting values SVD finds a μ-only plaid layer.

Rescale singular vectors to start ρ and κ

Backfitting Given $\rho_{ik}, \kappa_{jk} \in \{0, 1\}$ for $k = 1, \ldots, K$ it is cheap to re-estimate all the θ_{ij}.

Choosing K Permute row contents, then columns.

Stop if the algorithm finds more structure in the permuted data. Negative binomial regularization.

Stepping Use ≈ 10 steps to get ρ_i, κ_j into $\{0, 1\}$.

Unisign We may want a common sign for $\mu + \alpha_i$.

Robustness Inspect each new found layer: release any rows or columns not well explained.

Food data

$n = 961$ foods

$p = 6$ measures:

1. Fat proportion
2. Saturated fat proportion
3. Calories per gram
4. Cholesterol proportion $\times 1000$
5. Protein proportion
6. Carbohydrate proportion

For each column: subtract mean, divide by st.dev.

Source:

http://www.ntwrks.com/∼mikev/chart1.html

Yeast data

Name	Samples
Alpha	1–18
Elutriation	19–32
CDC	33-47
Sporulation	48–53
Sporulation-5	54–56
Sporulation-	57–58
Heat Shock	59–64
DTT	65–68
Cold	69–72
Diauxic Shift	73–79

Eisen, Spellman, Brown, Botstein: PNAS (1998)

Data Analysis

- Analyze log expression
- Few missing values: imputed by additive model
- Background Layer
  - Full model $\mu + \alpha_i + \beta_j$
  - All genes $\rho_i = 1, \forall i$
  - All samples $\kappa_j = 1, \forall j$
- Mine the interaction, with up to 40 layers
  - unisign
  - 50% threshold
  - 3 permutations per round
- Search stopped at 34 layers: 35th had zero genes

Future directions

- Refine existing algorithm
- Explore information retrieval applications
- Explore recommender system applications
- Find less greedy version
- Incorporate predictors
- Extend to higher way tables of data
- Larger data sets (99% missing)
- Use covariates
- Are there “plaid-lets”?

Code

Available for academic research

www-stat.stanford.edu/∼owen/clickwrap/plaid.html

Refinements

Replace

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_i\sum_j \left(Z_{ij} - \rho_i\kappa_j\left[\mu + \alpha_i + \beta_j\right]\right)^2 \\ \text{s.t.} \quad 0 &= \sum_i \rho_i^2\alpha_i = \sum_j \kappa_j^2\beta_j \\ \rho_i, \kappa_j &\in \mathbb{R} \end{aligned}$$

By:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_i\sum_j \left(Z_{ij} - \left[\mu + \alpha_i + \beta_j\right]\right)^2 \\ & + \frac{1}{2}\sum_i\sum_j Z_{ij}^2(1 - \rho_i\kappa_j) \\ \text{s.t.} \quad 0 &= \sum_i \rho_i\alpha_i = \sum_j \kappa_j\beta_j \\ \rho_i, \kappa_j &\in [0,1] \end{aligned}$$

Updates become:

Model parts

$$\mu \leftarrow \frac{\sum_i \sum_j \rho_i \kappa_j Z_{ij}}{\sum_i \sum_j \rho_i \kappa_j}$$

$$\alpha_i \leftarrow \frac{\sum_j \kappa_j (Z_{ij} - \mu)}{\sum_j \kappa_j}$$

$$\beta_j \leftarrow \frac{\sum_i \rho_i (Z_{ij} - \mu)}{\sum_i \rho_i}$$

Memberships

$$\rho_i \leftarrow 1 \text{ iff } \sum_j \kappa_j \left[(Z_{ij} - \theta_{ij})^2 - Z_{ij}^2 \right] < 0$$

$$\kappa_i \leftarrow 1 \text{ iff } \sum_i \rho_i \left[(Z_{ij} - \theta_{ij})^2 - Z_{ij}^2 \right] < 0$$

So far:

1. Seems to find slightly better layers
2. Harder to frame multi-layer model

Another refinement

Optimize “Mean Square” instead of “Sum of Squares”

Gets smaller more intense layers

Individual layers more interpretable

But more of them required

Requires a tradeoff of intensity vs sum of squares