Dimension reduction for microarray data using latent variables

Jelle Goeman
Sara van de Geer
Hans van Houwelingen

Leiden University Medical Center &
Leiden University Mathematical Institute

Workshop on Statistical Analysis of Gene Expression Data, Wye, 12-7-2003

Introduction: Goals

- Goals of microarray prediction methods:

- Primary goal:
  To predict the outcome y for a new individual
  using his/her array data vector x

- Secondary goal:
  To interpret the prediction rule in terms of genes
  $\rightarrow$ which genes play a large role?

- Problem:
  Genes may be missing from the new array

Regression models

- Usual choice in prediction modelling:
- Make a linear model for $y|x_1, \ldots, x_p$
- And use dimension reduction techniques because $p \gg n$
  - Variable selection
  - Ridge Regression
  - Principal Components Regression
  - Partial least squares
  - ...
- All work OK: good results for prediction

But: Drawbacks of regression models

- Regression coefficient for gene i is not a quality of gene i
  adding/romoving genes changes coefficients of other genes
  $\rightarrow$ interpretation not trivial
  $\rightarrow$ problem if genes are missing in the new array

- Much biological & technical information not used:
  - Expression measurements are inaccurate $\rightarrow$ noise
  - Many genes are highly correlated
    because genes react to each other
    and genes are functionally related (pathways)

Solution: Joint modelling

- Model for joint distribution of $x_1, \ldots, x_p$ and x
- Basic idea: underlying latent structure

$$\begin{tikzcd}
x_1, \ldots, x_p & & y \\
& f_1, \ldots, f_m \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

- $f_1, \ldots, f_m$ is the latent 'biological state'
- Dimension of the biological problem is $m \ll p$
- Motivation:
  - Biological: cells have limited number of functions
  - Statistical: dimension reduction techniques are successful

Advantages and disadvantages of joint modelling

- Parameters do not change if genes are added/removed
- Error in $x_1, \ldots, x_p$ explicitly modelled
- Yields a biologically interpretable structure (hopefully)
- Model is multi-functional:
  - Use for prediction
    $\rightarrow$ look at $y|x_1, \ldots, x_p$
  - Use for finding differential expression
    $\rightarrow$ look at $x_i|y$
  - Clustering
- Disadvantage: complexity

The factor analysis model

- Suppose all relationships are linear:

$$\begin{aligned} x_i &= \mu_i + f'\alpha_i + \varepsilon_i \\ y &= \mu_y + f'\beta + \varepsilon_y \end{aligned}$$

- Suppose y and x independent given $f = (f_1, \ldots, f_m)'$ with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)' \sim \mathcal{N}(0, \Psi)$ and $\varepsilon_y \sim \mathcal{N}(0, \sigma_y^2)$

- Suppose Ψ well-conditioned (e.g. diagonal)

- Suppose $f \sim \mathcal{N}(0, I_m)$

- $\approx$ normal linear factor analysis model

The joint distribution

- Notation: $A = (\alpha_1, \ldots, \alpha_p)$, an $m \times p$ matrix of 'loadings': The coefficients of the genes w.r.t. the latent state

- The joint distribution of $(y, x_1, \ldots, x_p)'$ is normal

- With mean $(\mu_y, \mu_1, \ldots, \mu_p)'$ and covariance matrix

$$\begin{pmatrix} \beta'\beta + \sigma_y^2 & \beta' A \\ A'\beta & A'A + \Psi \end{pmatrix}$$

- Rank m matrix + well-conditioned matrix

- Note: model is over-parameterized

The blessings of dimensionality

- Special thing about microarray data is large p
  $\rightarrow$ use estimation procedures that are good for $p \rightarrow \infty$

- Let Θ be well-conditioned $p \times p$ positive definite
  Then $X\Theta X'$ is $n \times n$ weighted 'covariance' matrix

- Let P be projection matrix for projection
  on the first $q \leq m$ eigenvalues of $X\Theta X'$

- As $p \rightarrow \infty$ P projects on subspace of row space of F
  where F is the $m \times n$ matrix of all latent variables

- For large p we can act almost as if F known

Consequences

- In normal factor analysis:
  - Estimate A given Ψ by maximum likelihood
  - Estimate Ψ given A by maximum likelihood
  - Maximum Likelihood breaks down if $p \geq n$
- If $p \gg n$:
  - Estimate A given Ψ by maximum likelihood
    but: estimate does not depend on Ψ if $p \to \infty$
  - Estimate Ψ by method-of-moments
    estimate does not depend on A
  - Stable: no iteration needed

Prediction

- $\mathrm{E}(y|x)$ linear in x

$$\mathrm{E}(y|x) = \mu_y + \gamma' x$$

- Regression coefficients:

$$\begin{aligned} \gamma &= (A'A + \Psi)^{-1} A'\beta \\ &= \Psi^{-1} A'(A\Psi^{-1}A' + I_m)^{-1}\beta \end{aligned}$$

- Problem: Ψ unknown and hard to estimate

Alternative regression coefficients

- Use estimated Ψ in

$$\gamma = \Psi^{-1} A' (A \Psi^{-1} A' + I_m)^{-1} \beta$$

- Gives large variance if $\hat{\Psi}$ is near-singular
- Alternative with fixed Θ well-conditioned

$$\gamma_\Theta = \Theta^{-1} A' (A \Theta^{-1} A' + I_m)^{-1} \beta$$

- Gives good predictions for a wide range of Θ
- Usually better than $\hat{\Psi}$ unless n large
- Alternative: $\hat{\Psi}$ shrunk to become better conditioned

Prediction error

- Prediction error for predicting y_{n+1} for a new array x_{n+1}
- But: distribution of x_{n+1} can be estimated
- Calculate prediction error averaged over possible new x_{n+1}
- Main sources of prediction error:
  - Unpredictable part of y_{n+1}
  - Error from overfit of y
  - Error from noise in x_{n+1}
  - Error from overfit of x
- We have derived explicit formula's for these (valid for large p)

Comparison with Principal Components Regression

- Close connection to principal components regression
- Take $\Theta = \sigma^2 I_p$ and let $\sigma^2 \to 0$
- γ_Θ becomes principal components regression coefficients
- Differences:
  - $\sigma^2 > 0$ also shrinks first few components
  - $\Psi \neq \sigma^2 I_p$ allows weighting of genes with prior information
  - Using information in $\hat{\Psi}$ allows sifting of unreliable genes

Comparison with Ridge Regression

- Also connection with ridge regression
  - Take $\Theta = \sigma^2 I_p$
  - Assume $m \geq n$
- Gives shrinkage similar to ridge regression
- σ^2 takes place of coefficient of penalty:
- Slightly different shrinkage function
- Also connections to partial least squares and variable selection

Discussion

- Interpretable prediction model much harder than prediction model
- But: doable
- Resulting model still quite general and intuitive
- Close connections to existing methods
- Nice mathematical structure
- Explicit assumptions
- Possiblilty to incorporate technical/biological information