Variable Selection for High-Dimensional Data with Spatial-Temporal Effects and Extensions to Multitask Regression and Multicategory Classification

Tong Tong Wu

Department of Epidemiology and Biostatistics
School of Public Health
University of Maryland, College Park

October 31, 2009

Outline

Overview and Introduction

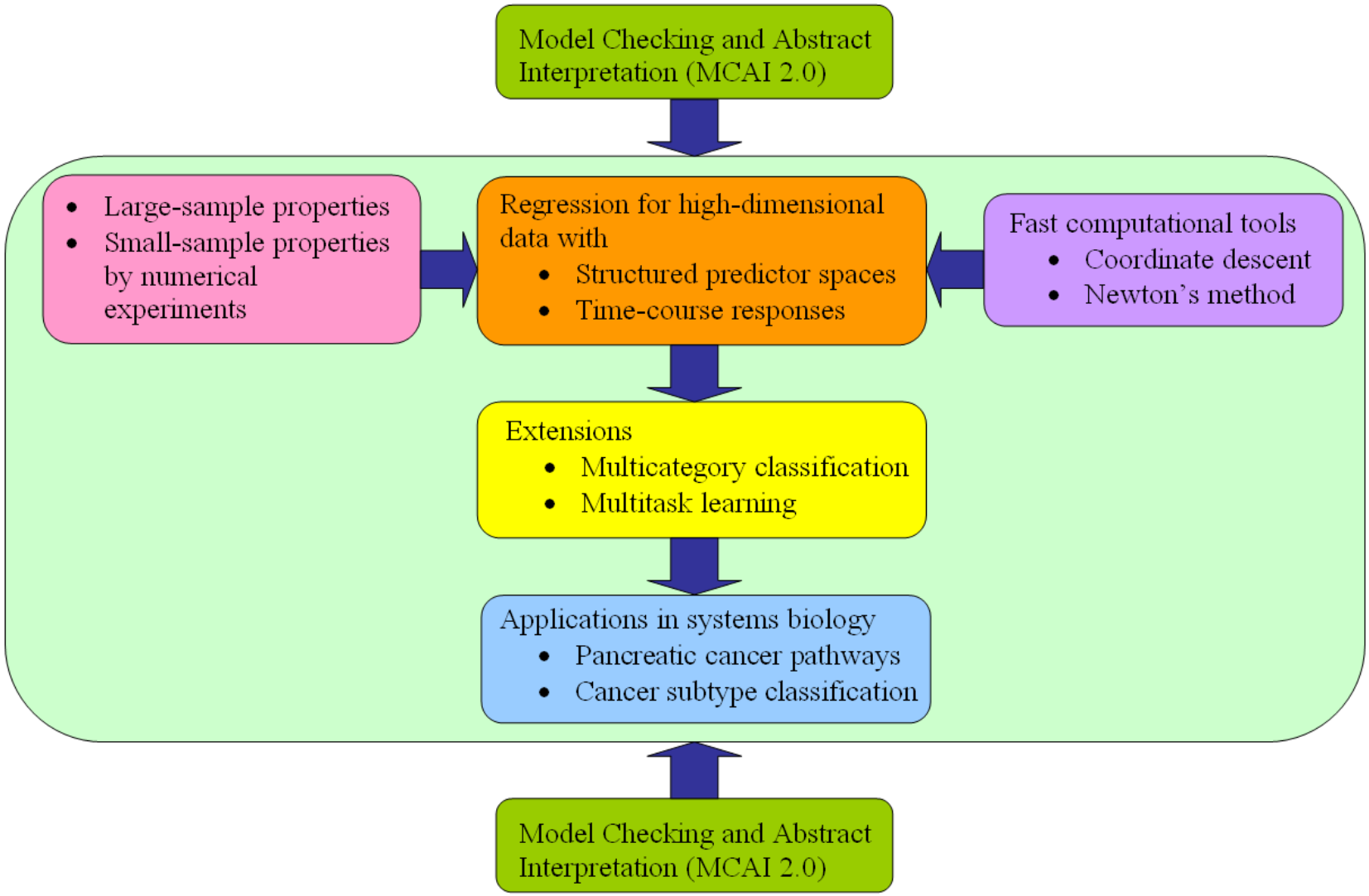


Figure 1: Overview of the proposed research

Objectives: NSA Mathematical Science Program

1. Development of new variable selection method for high-dimensional data with spatial-temporal effects in regression
    1. Models for time-course data (temporal effects): AR(q), kernel combined regression models
    2. New penalty functions for structured predictor spaces (spatial effects): linear chain, undirected graph
    3. Fast and stable algorithms for high-dimensional spatial-temporal data
    4. Asymptotic properties
    5. Finite sample properties by numerical experiments
2. Extensions to multicategory classification and multitask regression
3. Applications in biology: pancreatic cancer pathways, cancer subtype classification

Variable Selection

- Design of effective and efficient tools to extract scientific insights from these massive and noisy high-throughput data
- Two approaches for dimension reduction: *feature selection* and *feature extraction*
- Selection of a possible best subset from the original variables to build robust learning models
- Goals of variable selection
  - Alleviating the effect of the curse of dimensionality
  - Acquiring better understanding about data
  - Enhancing generalization capability and prediction
  - Improving stability
  - Avoiding bias in hypothesis tests during or after variable selection
  - Retaining scientific interpretability
  - Speeding up learning process

Lasso Penalized Regression

Lasso (Least absolute shrinkage and selection operator) penalized regression (Tibshirani 1996) can be phrased as

$$\min \quad f(\theta) \quad = \quad g(\theta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

- $g(\theta)$: loss function
  - $g(\theta) = \sum_{i=1}^{n} |y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j|$ for ℓ_1 regression
  - $g(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ for ℓ_2 regression
- $\theta = (\mu, \beta_1, \ldots, \beta_p)^t$: parameter vector
- $\lambda \sum_{j=1}^{p} |\beta_j|$: lasso penalty
- λ: positive tuning constant that determines the strength of the lasso penalty

Lasso Penalty $\lambda \sum_{j=1}^{p} |\beta_j|$

- Shrinking each β_j toward the origin
- Discouraging models with large numbers of marginally relevant predictors
- More effective in deleting irrelevant predictors than a ridge penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ because $|b| > b^2$ for small b

Lasso vs. Ridge

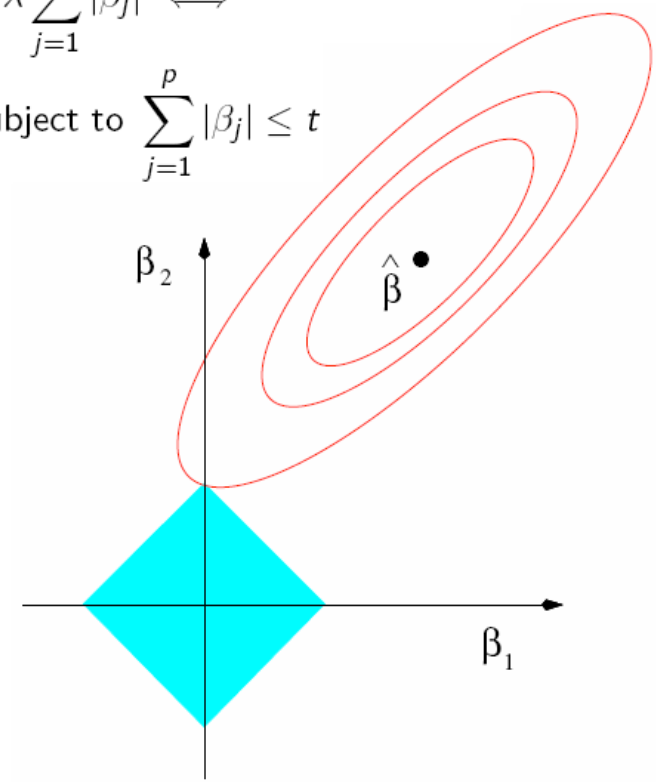
In ℓ_2 regression: $\sum_{i=1}^{n}(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0) + c$

Lasso vs. Ridge

In ℓ_2 regression: $\sum_{i=1}^{n}(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0) + c$

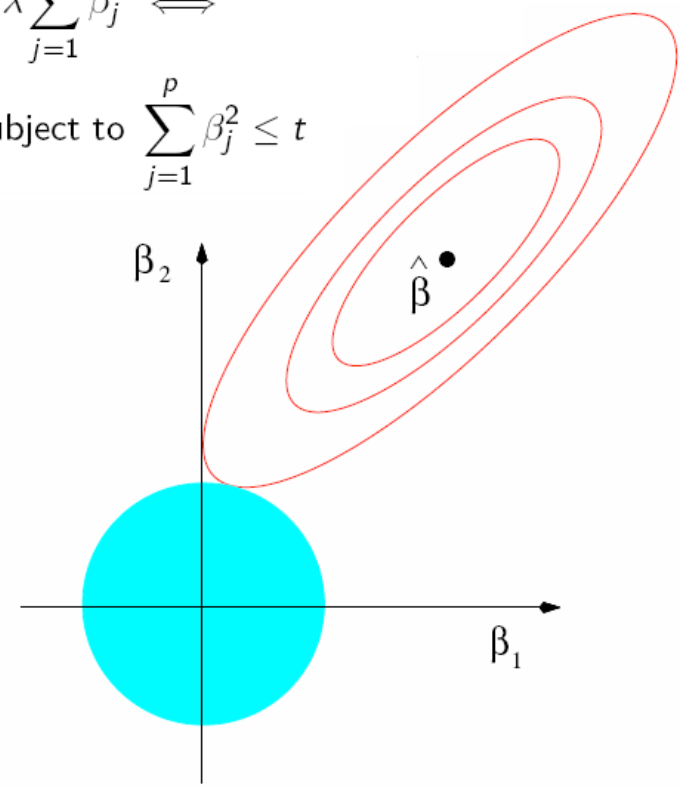
$$\min\ g(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| \iff$$

$$\min\ g(\theta) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le t$$



$$\min\ g(\theta) + \lambda \sum_{j=1}^{p} \beta_j^2 \iff$$

$$\min\ g(\theta) \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \le t$$



Variable Selection for High-Dimensional Data with Spatial-Temporal Effects

Spatial-Temporal Models

Idea: Minimize a regularized objective function

$$\min \quad f(\theta) \quad = \quad g(\theta) + \lambda P(\beta)$$

- $g(\theta)$: loss function for time-course model (temporal effects)
- $P(\beta)$: penalty function for selecting structured predictors (spatial effects)

Regression Models for Time-Course Data

- Regression model for subject i at time t_s, $s = 1, \ldots, T$

$$y_i(t_s) \quad = \quad \mu(t_s) + x_i^t \beta(t_s) + \epsilon_i(t_s)$$

where

- $y_i(t)$: responses at time t
- $x_i = (x_{i1}, \ldots, x_{ip})^t$: fixed p predictors
- $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))^t$: time-varying coefficients
- $\mu(t)$: time-varying intercept

Regression Models for Time-Course Data

- Regression model for subject i:

$$\begin{pmatrix} y_i(t_1) \\ \vdots \\ y_i(t_T) \end{pmatrix} = \begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_T) \end{pmatrix} + \begin{pmatrix} \beta_1(t_1) & \dots & \beta_p(t_1) \\ \vdots & & \\ \beta_1(t_T) & \dots & \beta_p(t_T) \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} + \begin{pmatrix} \epsilon_i(t_1) \\ \vdots \\ \epsilon_i(t_T) \end{pmatrix}$$

$$y_i \equiv \mu + Bx_i + \epsilon_i$$

  where B is a $T \times p$ coefficient matrix whose sth row is $\beta(t_s)^t$

- Linear system for n observations:

$$\begin{pmatrix} y_1^t \\ \vdots \\ y_n^t \end{pmatrix} = \mathbf{1}\mu^t + \begin{pmatrix} x_i^t \\ \vdots \\ x_n^t \end{pmatrix} \begin{pmatrix} \beta_1(t_1) & \dots & \beta_p(t_1) \\ \vdots & & \\ \beta_1(t_T) & \dots & \beta_p(t_T) \end{pmatrix}^t + \begin{pmatrix} \epsilon_1^t \\ \vdots \\ \epsilon_n^t \end{pmatrix}$$

Regression Models for Time-Course Data

- Stationary process

$$\mathrm{E}\left[\epsilon(t)\right] = 0 \quad \text{and} \quad \mathrm{Cov}(\epsilon_t, \epsilon_{t+s}) = \mathrm{Cov}(\epsilon_t, \epsilon_{t-s}) = \sigma^2 \rho_s$$

  where ρ_s is the error autocorrelation at lag s

- Error covariance matrix Σ for each subject

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{T-3} \\ \vdots & & & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma^2 V$$

- Loglikelihood

$$-\frac{n}{2}\log(\det\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu - Bx_i)^t \Sigma^{-1}(y_i - \mu - Bx_i)$$

AR(q) Model

- Autocorrelated regression errors ϵ_t

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} + v_t,$$

  where the "random shocks" $v_t \sim N(0, \sigma_v^2)$
- Given that first q observations for each subject are fixed, the conditional loglikelihood of the remaining $T - q$ observations $(y(t_{q+1}), \ldots, y(t_T))$

$$-\frac{n}{2}\log\sigma_v^2 - \frac{1}{2\sigma_v^2}\underbrace{\sum_{i=1}^{n}\sum_{s=q+1}^{T}\left\{y_i(t_s) - \mu(t_s) - x_i^t\beta(t_s) - \sum_{j=1}^{q}\phi_j\left[y_i(t_{s-j}) - \mu(t_s) - x_i^t\beta(t_{s-j})\right]\right\}^2}$$

- Minimizing the least-squares type loss function

$$g_1(\theta) = \sum_{i=1}^{n}\sum_{s=q+1}^{T}\left\{y_i(t_s) - \mu(t_s) - x_i^t\beta(t_s) - \sum_{j=1}^{q}\phi_j\left[y_i(t_{s-j}) - \mu(t_s) - x_i^t\beta(t_{s-j})\right]\right\}^2$$

Kernel Smoothed Regression Model

- Minimizing the kernel smoothed loss function at time t_s for $\theta(t_s)$

$$\hat{\theta}(t_s) = \arg\min \sum_{i=1}^{n} \sum_{r=1}^{T} w(t_s, t_r) \Big\{ y_i(t_r) - \mu(t_s) - x_i^t \beta(t_s) \Big\}^2$$

  where $\sum_{r=1}^{T} w(t_s, t_r) = 1$
- Kernel smoothed response

$$\tilde{y}_i(t_s) = \sum_{r=1}^{T} w(t_s, t_r) y_i(t_r)$$

- Minimizing the least-squares type loss function

$$g_2(\theta) = \sum_{i=1}^{n} \sum_{s=1}^{T} \Big\{ \tilde{y}_i(t_s) - \mu(t_s) - x_i^t \beta(t_s) \Big\}^2$$

Structured Predictor Spaces

- **Grouped predictors** (Wu and Lange 2008; Yuan and Lin 2006; Zhao et al. 2009; Zou and Hastie 2005)
  Example: in gene microarray experiments, genes can sometimes be grouped into biochemical pathways subject to genetic coregulation and their expression levels are expected to be highly correlated
- **Order restrictions** (Efron et al. 2004; Turlach et al. 2005; Wu et al. 2009; Yuan et al. 2007)
  Example: a higher order term should be selected only when the corresponding lower order terms are present in the model
- **Ordered predictors** (Tibshirani et al. 2005; Li and Zhang 2008)
  Example: genes or SNPs are located on chromosomes as a sequence

Structured Predictor Spaces

- Important to incorporate the structural information into the model building process when the covariate space is highly structured
- Two structures of predictor spaces
  - Linear chain: sequencing and mapping information of genomes
  - Undirected graph: networking information

Linear Chain Structure

- A genome map describes the order of genes or other markers and the spacing between them
- The existing works incorporating linear chain structures only consider the sequencing of markers (e.g. fused LASSO (Tibshirani et al. 2005), Bayesian variable selection (Li and Zhang 2008; Tai and Pan 2009))
- The distance between two markers, which provides information on the recombination of the two markers and their participation together in relevant biological processes, however, has not been taken into consideration
- The sequencing and mapping information is widely available for public use in the internet

Penalty Function for Linear Chain Structure

- Adaptive fused lasso penalty

$$P_1(\beta) = \gamma \sum_{j=2}^{p} \frac{1}{|m_j - m_{j-1}|} |\beta_j - \beta_{j-1}| \equiv \sum_{j=2}^{p} \gamma_j |\beta_j - \beta_{j-1}|,$$

where m_j is the map distance of marker j, and
$\gamma_j = \gamma / |m_j - m_{j-1}|$ is the adaptive weight

Undirected Graphical Structure

- Example: Nodes in the graph represent genes and edges represent interactions between genes
- Bayesian approach: Li and Zhang (2008); Tai and Pan (2009)
- Frequentist approach: Li and Li (2008); Pan et al. (2009)
- Employing a different penalty function to incorporate the mapping information for a more straightforward biological reason

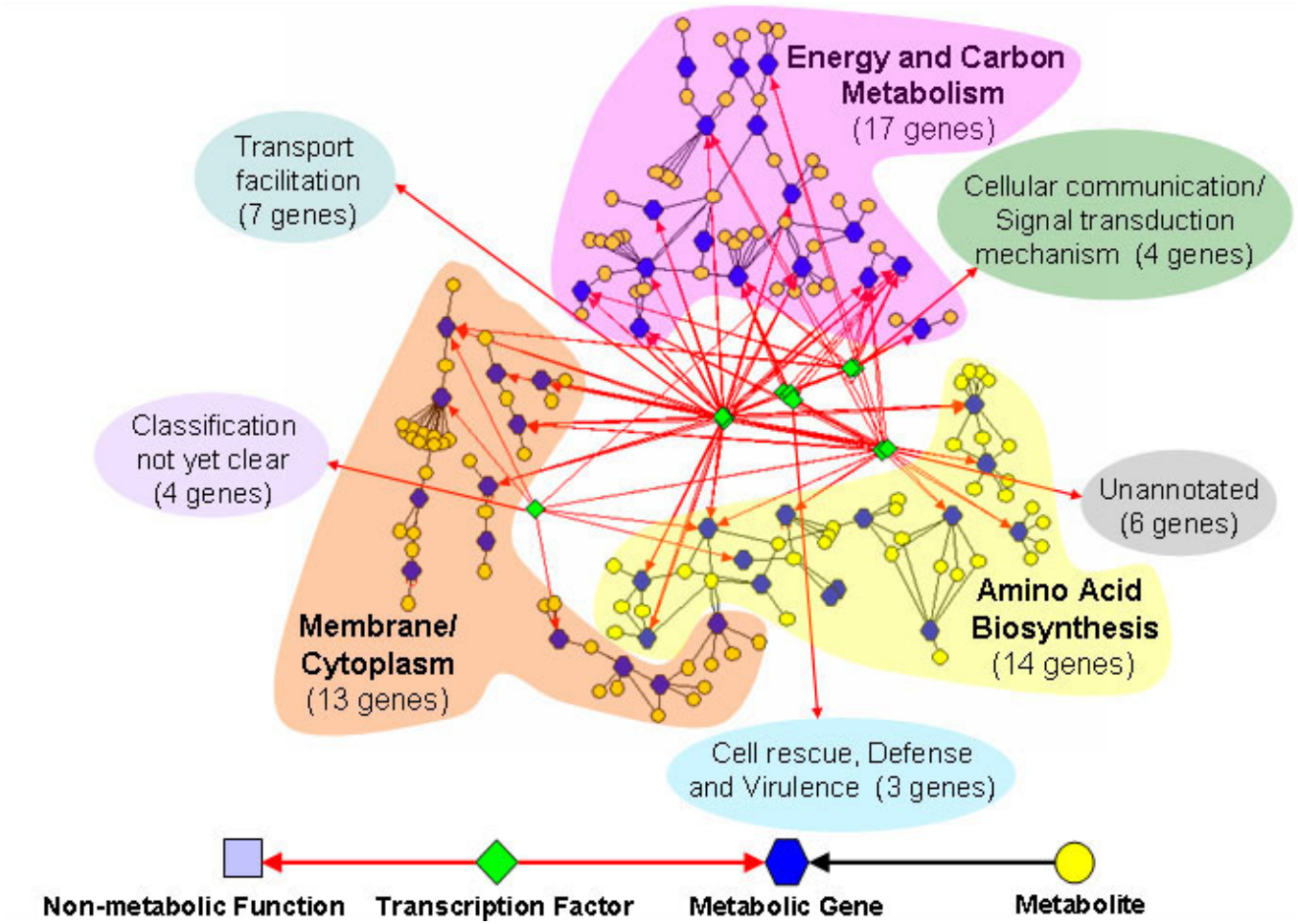


Figure 2: Gene regulatory networks. Each node is a single gene, and any two genes are connected if in the same pathway (reproduced with permission from Sinauer Associates, Inc. (Gifford et al. 2006))

Penalty Function for Undirected Graphical Structure

- Consider a weighted graph $G = (V, E, W)$, where
  - V: the set of nodes corresponding to the p predictors
  - $E = \{u \sim v\}$: the set of edges indicating whether the predictors u and v are linked in the network
  - W: the weights of the edges with $w(u, v)$ denoting the weight of edge $e = (u \sim v)$
- In this study, the edge weights represent the map distances between nodes
- Penalty for graphical structure:

$$P_2(\beta) = \gamma \sum_i \sum_{j \sim i} w(i,j)|\beta_i - \beta_j| \equiv \sum_i \sum_{j \sim i} \gamma_{ij}|\beta_i - \beta_j|$$

where $\gamma_{ij} = \gamma w(i,j) = \gamma / |m_i - m_j|$

Spatial-Temporal Models

- The spatial-temporal model therefore can be implemented by minimizing the objective function

$$f(\theta) = g_k(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| + \gamma_l P_l(\beta), \quad k, l = 1, 2 \qquad (1)$$

- The two ℓ_1 penalties encourage sparsity in the coefficients and their differences, respectively
- The objective function is convex but nondifferentiable, with a large number of predictors
- Use the cyclic coordinate descent algorithm highlighted in Wu and Lange (2008) and Friedman et al. (2007)
- Allow response to be related to a different subset of predictors at each time point

Literature Review: Computational Algorithms

The inefficiency of the original lasso limited its impact on statistical practice (Madigan and Ridgeway 2004), even with the more recent algorithms (Osborne et al. 2000) due to their complexity

Literature Review: Computational Algorithms

The inefficiency of the original lasso limited its impact on statistical practice (Madigan and Ridgeway 2004), even with the more recent algorithms (Osborne et al. 2000) due to their complexity

- Efron et al. (2004): LARS (Least Angle Regression, Lasso, and Forward Stagewise)
- Fu (1998) and Daubechies et al. (2004): coordinate descent for lasso penalized ℓ_2 regression, but no follow up for underdetermined problems
- Wang et al. (2006): solution path for penalized ℓ_1 regression using linear programming
- Park and Hastie (2006): solution path for penalized ℓ_2 regression and generalized linear models
- Wu and Lange (2008) and Friedman et al. (2007): independent and concurrent work on coordinate descent!

Coordinate Descent Algorithms

- The difficulties of minimizing the objective function (1) lie in three facts
  1. Nondifferentiability due to the ℓ_1 penalties
  2. $p \gg n$
  3. Impossibility of matrix operations for huge p
- Coordinate descent: simplicity, speed, and stability (Friedman et al. 2007; Wu and Lange 2008; Wu et al. 2009)
- Decisive advantages in dealing with data sets in our setting
- Standard version of coordinate descent: cycling through the parameters and updating each in turn

Coordinate Descent Algorithms

Directional derivatives along forward or backward directions, e.g. e_k is the coordinate direction along which β_k varies

- Forward directional derivative of the objective function

$$d_{e_k} f(\theta) = \lim_{\tau \downarrow 0} \frac{f(\theta + \tau e_k) - f(\theta)}{\tau} = d_{e_k} g(\theta) + \begin{cases} \lambda & \beta_k \geq 0 \\ -\lambda & \beta_k < 0 \end{cases}$$

- Backward directional derivative of the objective function

$$d_{-e_k} f(\theta) = \lim_{\tau \downarrow 0} \frac{f(\theta - \tau e_k) - f(\theta)}{\tau} = d_{-e_k} g(\theta) + \begin{cases} -\lambda & \beta_k > 0 \\ \lambda & \beta_k \leq 0 \end{cases}$$

Procedures

1. Start all parameters at the origin
2. Evaluate both $d_{e_k} f(\theta)$ and $d_{-e_k} f(\theta)$ for β_k
   - If both are nonnegative $\Rightarrow$ skip the update for β_k
   - If either directional derivative is negative $\Rightarrow$ solve for the minimum in that direction
   - Both directional derivatives cannot be negative because this contradicts the convexity of $f(\theta)$s
3. After the direction is chosen, Newton's method can be used for updating the parameter
4. Check if the objective function is driven downhill
5. Halve the step size if the descent property fails

Underdetermined Problems

$p \gg n$ with just a few relevant predictors

- Fast speed of cyclic coordinate descent
  - Most updates are skipped and the corresponding parameters never budge from their starting values of 0
  - Complete absence of matrix operations
- Numerical stability from the descent property of each update

Extensions to Multicategory Classification and Multitask Regression

Overview of Vertex Discriminant Analysis (VDA)

- A new method of supervised learning
- Based on linear discrimination among the vertices of a regular simplex in Euclidean space
- Each vertex represents a different category
- A regression problem involving ϵ-insensitive residuals and penalties on the coefficients of the linear predictors
  - Ridge penalty (VDA_R)
  - Lasso penalty (VDA_L)
  - Euclidean penalty (VDA_E, VDA_LE)
- Minimizing the objective function by a primal MM algorithm or a coordinate descent algorithm
- Competitive in statistical accuracy and computational speed

Regularized ϵ-insensitive Loss Function for VDA

VDA (Vertex Discriminant Analysis) discriminates among k categories by minimizing ϵ-insensitive loss plus penalties

$$f(\theta) = \sum_{i=1}^{n} g(y_i - Bx_i - \mu) + \lambda_L \sum_{j=1}^{k-1} \sum_{l=1}^{p} |\beta_{jl}| + \lambda_E \sum_{l=1}^{p} \|\beta_{(l)}\|_2$$

where

- $y_i \in R^{k-1}$ is the vertex assignment for case i
- $\beta_{(l)}$ is the lth column of a $k \times p$ matrix B of regression coefficients
- μ is a $(k-1) \times 1$ column vector of intercepts
- $g(v)$: modified ϵ-insensitive loss

$$g(v) = \begin{cases} \|v\|_2 - \epsilon & \text{if } \|v\|_2 \geq \epsilon + \delta \\ \frac{(\|v\|_2 - \epsilon + \delta)^3 (3\delta - \|v\|_2 + \epsilon)}{16\delta^3} & \text{if } \|v\|_2 \in (\epsilon - \delta, \epsilon + \delta) \\ 0 & \text{if } \|v\|_2 \leq \epsilon - \delta \end{cases}$$

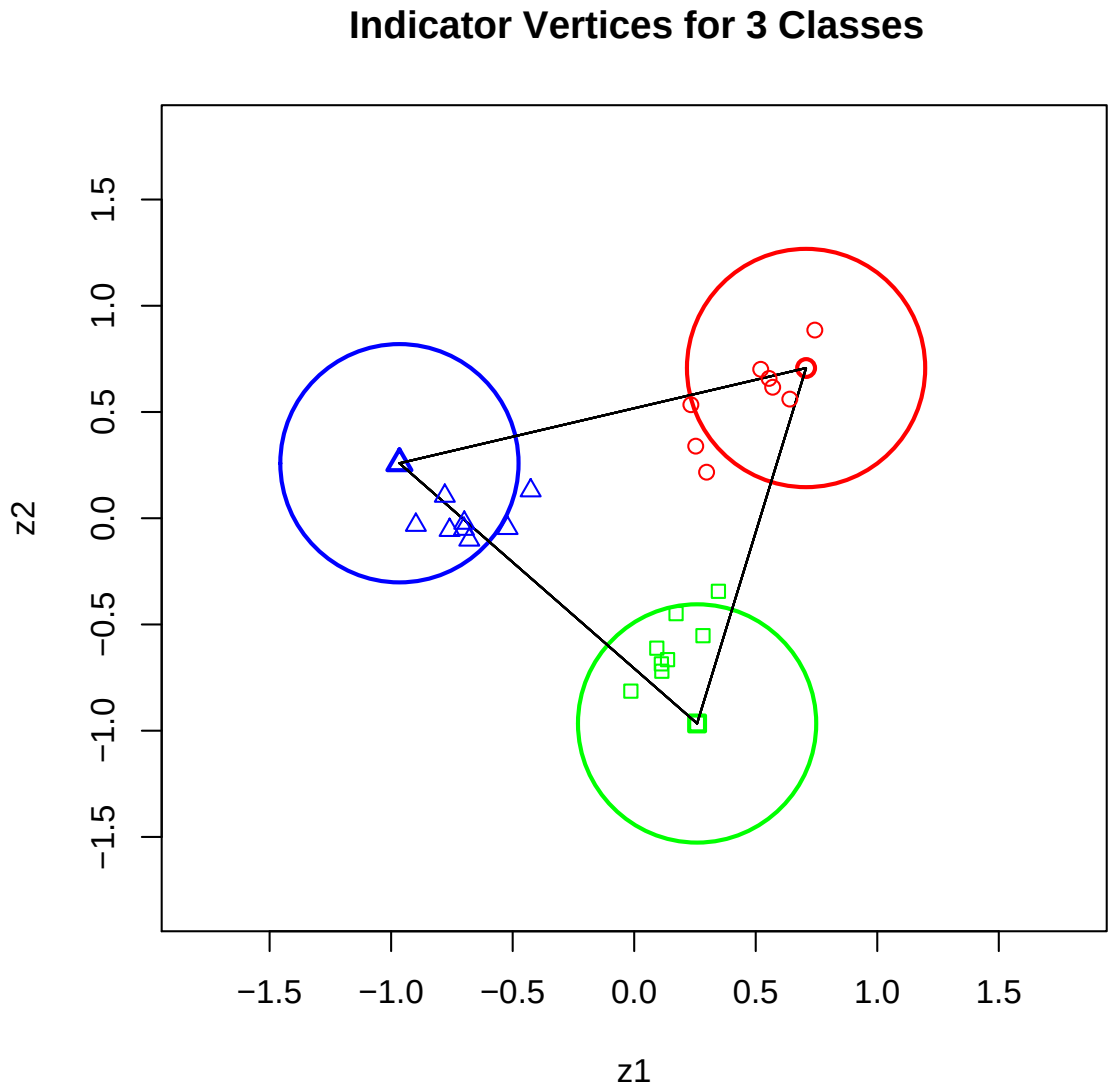


Figure 3: Plot of the indicator vertices for 3 classes, with a ϵ-radius circle around each vertex

VDA with Spatial Effects

- Spatial information can be incorporated into classification by adding the network-constraint penalty to the regularized objective function

$$f(\theta) = \sum_{i=1}^{n} g(y_i - Bx_i - \mu) + \lambda_L \sum_{j=1}^{k-1} \sum_{l=1}^{p} |\beta_{jl}| + \lambda_E \sum_{l=1}^{p} \|\beta_{(l)}\|_2 + \gamma_l P_l(A)$$

$l = 1, 2$

- Multiresponse regression problem: $y_i \in R^{k-1}$ for k categories
- Cyclic coordinate descent algorithm

Multitask Regression with Spatial Effects

- Learning multiple related tasks from data simultaneously can be advantageous to learning these tasks independently (Bakker and Heskes 2003; Caruana 1997; Evgeniou and Pontil 2004)
- Main task is trained with other related problems at the same time using a shared representation
- Example: 24 soil samples with measurements of 14 quantities at 770 different wavelengths (CMIS/CSIRO)

Multitask Regression with Spatial Effects

- Assume we have k responses and share the same set of covariates
- For subject i, $y_i = (y_{i1}, \ldots, y_{ik})^t$ and $x_i = (x_{i1}, \ldots, x_{ip})^t$
- Further assume each task comes from a different regression model of x's
- Multiresponse regression model

$$\begin{pmatrix} y_1^t \\ \vdots \\ y_n^t \end{pmatrix} = \mathbf{1}\mu^t + \begin{pmatrix} x_i^t \\ \vdots \\ x_n^t \end{pmatrix} \begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ & \ddots & \\ \beta_{k1} & \cdots & \beta_{kp} \end{pmatrix}^t$$

Discussion

Connections to MCAI 2.0

- Applications in biology
  1. Pancreatic cancer pathways
  2. Cancer subtype classification
- Model verification/checking by MACI 2.0

Thank you very much!

Bakker, B. and Heskes, T. (2003), "Task clustering and gating for Bayesian multitask learning," *Journal of Machine Learning Research*, 4, 83–99.

Caruana, R. (1997), "Multitask Learning," *Machine Learning*, 28, 41–75.

Daubechies, I., Defrise, M., and De Mol, C. (2004), "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, 57, 1413–1457.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Ann Stat*, 32, 407–499.

Evgeniou, T. and Pontil, M. (2004), "Regularized multitask learning," *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 109–117.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007), "Pathwise coordinate optimization," *Anns Appl Stat*.

Fu, W. J. (1998), "Penalized regressions: the bridge versus the lasso," *J Comp and Graph Stat*, 7, 397–416.

Gifford, M. L., Gutierrez, R. A., and Coruzzi, G. M. (2006), "Essay 12.2: Modeling the Virtual Plant: A Systems Approach to Nitrogen-Regulatory Gene Networks. Plant Physiology, Fourth Edition Online," online.

Li, C. and Li, H. (2008), "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, 24, 1175–1182.

Li, F. and Zhang, N. R. (2008), "Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics," *manuscript*.

Madigan, D. and Ridgeway, G. (2004), "Discussion of "least angle regression" by Efron et al." *Annals of Statistics*, 32, 465–469.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, 20, 389–403.

Pan, W., Xie, B., and Shen, X. (2009), "Incorporating Predictor Network in Penalized Regression with Application to Microarray Data," *Biometrics*, online publication.

Park, M. Y. and Hastie, T. (2006), "Penalized logistic regression for detecting gene interactions," Tech. Rep. 2006-15, Department of Statistics, Stanford University.

Tai, F. and Pan, W. (2009), "Bayesian Variable Selection in Regression with Networked Predictors," *manuscript*.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *J. Roy. Stat. Soc., Series B*, 67, 91–108.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), "Simultaneous variable selection," *Technometrics*, 47, 349–363.

Wang, L., Gordon, M. D., and Zhu, J. (2006), "Regularized least absolute deviations regression and an efficient algorithm for parameter tuning," *Proceedings of the Sixth International Conference on Data Mining (ICDM'06). IEEE Computer Society*, 690–700.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009), "Genomewide association analysis by lasso penalized logistic regression," *Bioinformatics*, 25, 714–721.

Wu, T. T. and Lange, K. (2008), "Coordinate descent algorithms for lasso penalized regression," *Ann. Appl. Stat.*, 2, 224–244.

Yuan, M., Joseph, R., and Lin, Y. (2007), "An efficient variable selection approach for analyzing designed experiments," *Technometrics*, 49, 430–439.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *J Roy Stat Soc, Series B*, 68, 49–67.

Zhao, P., Rocha, G., and Yu, B. (2009), "The composite absolute penalties family for grouped and hierarchical variable selection," *Annals of Statistics*, 37, 3468–3497.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *J Roy Stat Soc, Series B*, 67, 301–320.