Independence screening approaches for Cox models with high dimensionality

Manuela Zucknick, Thomas Hielscher, Axel Benner

German Cancer Research Centre, Heidelberg

ISCB Prague, 26 August 2009

Introduction

In clinical applications of gene expression microarray data:

- **Typical objective:**
  Develop a prognostic model (gene signature) for overall survival. Two goals:
  - Good prognostic value/ prediction accuracy
  - Biological interpretation (sparsity)
- **Possible solution:**
  Sparse Cox PH regression models by using penalised partial log-likelihood:

$$Q(\beta) = \sum_{i=1}^{n} \ell_i(\beta) - \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|)$$

Introduction

In clinical applications of gene expression microarray data:

- **Typical objective:**
  Develop a prognostic model (gene signature) for overall survival. Two goals:
  - Good prognostic value/ prediction accuracy
  - Biological interpretation (sparsity)
- **Possible solution:**
  Sparse Cox PH regression models by using penalised partial log-likelihood:

$$Q(\beta) = \sum_{i=1}^{n} \ell_i(\beta) - \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|)$$

Introduction

Aim: Combine good prediction performance with

- consistent model selection ($\lim_{n\to\infty} Pr(\hat{M}_n = M) = 1$),
- approximately unbiased parameter estimation.

- **Lasso** (Tibshirani 1996): model selection consistent but (downward) biased parameter estimates
- **SCAD** (Fan and Li 2001): both properties fulfilled
  - Problem: needs a pre-selection step to reduce the number of variables to $p_{ini} < n$

Also: for very high-dimensional data with $n << p$, model selection performance can be improved by initial selection step.

Independence screening

Aim in practice: Keep all important regressors (TPR = 1), while eliminating most noise variables (small FPR).

Sure screening property:

Keep all important regressors with probability tending to 1.

A simple proposal... with a big impact:

Simple univariate scores can fulfill the sure screening property under certain conditions (Fan and Lv, 2008)

Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 5, *pp.* 849–911

Sure independence screening for ultrahigh dimensional feature space

Independence screening

Aim in practice: Keep all important regressors (TPR = 1), while eliminating most noise variables (small FPR).

Sure screening property:
Keep all important regressors with probability tending to 1.

A simple proposal... with a big impact:
Simple univariate scores can fulfill the sure screening property under certain conditions (Fan and Lv, 2008)

Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 5, *pp.* 849–911

Sure independence screening for ultrahigh dimensional feature space

Independence screening

In the linear (least squares) model:

- **Sure independence screening** (SIS, Fan and Lv 2008)
  - Also called: correlation learning, component-wise regression
  - $\omega = X^T y$ vector of component-wise regression coefficients (X,y centered and scaled)
  - Keep the p_{ini} variables with the largest $|\omega_j|$ values $\rightarrow$ with strongest **marginal correlations** with response

  Equivalent to sorting by model deviances of component-wise regression models (vs. null model)

- **PC-simple algorithm** (partial faithfulness, Bühlmann et al. 2009): from marginal correlations to partial correlations

Independence screening

In the linear (least squares) model:

- **Sure independence screening** (SIS, Fan and Lv 2008)
  - Also called: correlation learning, component-wise regression
  - $\omega = X^T y$ vector of component-wise regression coefficients (X,y centered and scaled)
  - Keep the p_{ini} variables with the largest $|\omega_j|$ values $\rightarrow$ with strongest **marginal correlations** with response

  Equivalent to sorting by model deviances of component-wise regression models (vs. null model)

- **PC-simple algorithm** (partial faithfulness, Bühlmann et al. 2009): from marginal correlations to partial correlations

Iterative sure independence screening (ISIS)

Potential problems with component-wise approaches:

- Unimportant variables, which are highly correlated with important predictors, can have substantial marginal correlation with response
- Important predictors, which are marginally uncorrelated but jointly correlated with response, will not get selected

Solution (Fan and Lv 2008; Fan, Samworth and Wu 2009):

- Apply SIS + penalised likelihood method iteratively ($\rightarrow$ active set M_t after iteration t)
- In iterations $t > 1$, compute model deviances ($j = 1, \ldots, p$) for
  - full model: covariate set $\{X_j\} \cup M_{t-1}$
  - reduced model: covariate set M_{t-1}

Iterative sure independence screening (ISIS)

Potential problems with component-wise approaches:

- Unimportant variables, which are highly correlated with important predictors, can have substantial marginal correlation with response
- Important predictors, which are marginally uncorrelated but jointly correlated with response, will not get selected

Solution (Fan and Lv 2008; Fan, Samworth and Wu 2009):

- Apply SIS + penalised likelihood method iteratively ($\rightarrow$ active set M_t after iteration t)
- In iterations $t > 1$, compute model deviances ($j = 1, \ldots, p$) for
  - full model: covariate set $\{X_j\} \cup M_{t-1}$
  - reduced model: covariate set M_{t-1}

ISIS and lasso

- **Connection with lasso** (for linear least squares):
  When selecting only one variable in each iteration of ISIS, the ISIS solution is similar to L_2-boosting with component-wise linear least squares $\longrightarrow$ similar to lasso. (Bühlmann in discussion of Fan and Lv 2008)

- Another motivation for using lasso as an initial estimator
  - SCAD with local linear approximation (SCAD-LLA, Zou and Li 2008): Starting with zero vector as initial estimator ($\beta^0 = 0$) will result in lasso solution after first iteration (β^1)

Independence screening for Cox regression

- **Adapting the SIS idea: SIScox**
  - Sure screening property holds for GLM models under uniform convergence property (Fan and Song 2009)
  - But: Does the sure screening property holds for the partial likelihood implementation for Cox models?
- **Cox Univariate Shrinkage (CUS, Tibshirani 2009)**
  - Developed specifically for the Cox model
  - Even stronger independence assumption than in linear models: independence both conditionally on each risk set and marginally

Independence screening for Cox regression

- **Adapting the SIS idea: SIScox**
  - Sure screening property holds for GLM models under uniform convergence property (Fan and Song 2009)
  - But: Does the sure screening property holds for the partial likelihood implementation for Cox models?
- **Cox Univariate Shrinkage (CUS, Tibshirani 2009)**
  - Developed specifically for the Cox model
  - Even stronger independence assumption than in linear models: independence both conditionally on each risk set and marginally

Independence screening for Cox regression

- **SIScox:** maximiser of partial log-likelihoods of component-wise models $Q(\beta_j) = \sum_{i=1}^{n} \ell_i(\beta_j)$
- **CUS:** maximiser of penalised partial log-likelihood $Q(\beta) = \sum_{i=1}^{n} \ell_i(\beta) - \sum_{j=1}^{p} \lambda(|\beta_j|)$
  - Problem is equivalent to maximising $Q(\beta_j) = \sum_{i=1}^{n} \ell_i(\beta_j) - \lambda(|\beta_j|)$ under assumption of conditional and marginal independence

Note: $\hat{\beta}_j^{CUS} \neq 0 \Leftrightarrow \text{score}_j > \lambda$
$\rightarrow$ Screening by CUS is equivalent to screening by score statistics from univariate Cox models

An additional problem for Cox regression

- **Omitting covariates results in different baseline hazards.**
- $\rightarrow$ Contrary to linear models:
  Even under the independence assumption, component-wise marginal Cox models do not give the same results as joint Cox models!
- Hence, correlation screening can miss important variables even under independence.

(Keiding, Andersen and Klein 1997, Schmoor and Schumacher 1997, Gerds and Schumacher 2001,...

"[A] model should be adjusted for possibly influential covariates even if the correct structure of dependence between the survival time and the covariates or within the covariates is uncertain." (Gerds and Schumacher 2001)

Simulation setup

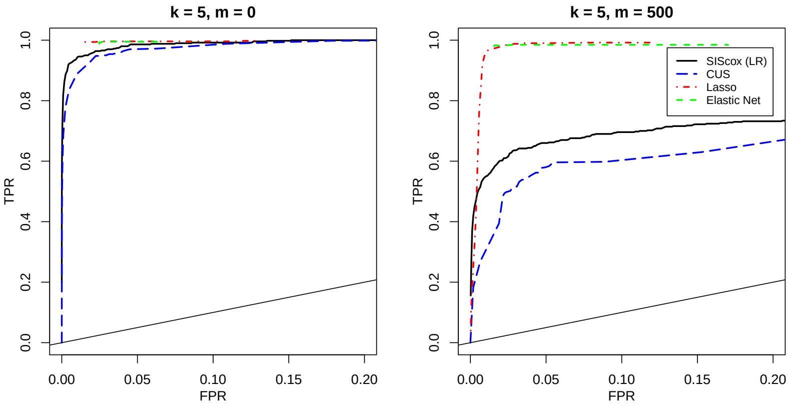
$B = 100$ **simulation runs with:**

- $p = 1000$ covariates; $n_1 = 200$ & $n_2 = 100$ (training & test)
- k randomly selected variables as predictors for survival response with $\beta_j = \log(2)$, for $p - k$ "noise" variables $\beta_j = 0$
- Pairwise correlations among the first m variables $\rho = 0.5$
- X_i marginally $\sim N(0, 1)$
- Survival times $\sim Exp(X\beta)$; censoring times $\sim Exp$ ($\approx 50\%$)

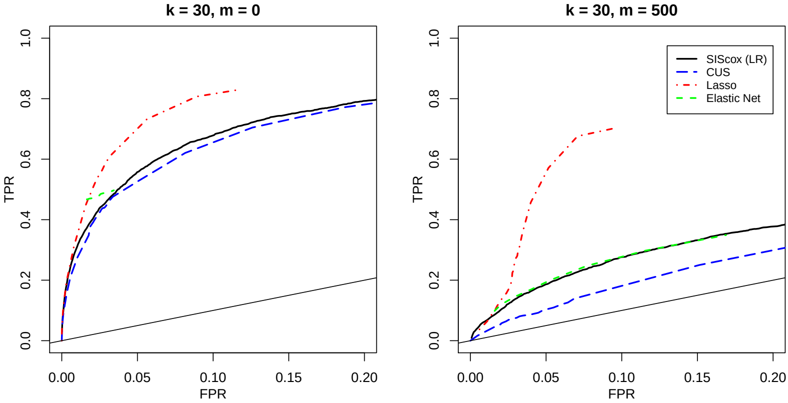
Scenarios:

- $k = 5,\ m = 0$
- $k = 5,\ m = 500$
- $k = 30,\ m = 0$
- $k = 30,\ m = 500$

Screening results (k = 5)



Screening results (k = 30)



Screening results for $p_{ini} = \lfloor n/\log(n) \rfloor = 37$

k = 5

	m = 0		m = 500	
	1-FPR	TPR	1-FPR	TPR
CUS	0.967	0.962	0.965	0.558
SIScox	0.968	0.970	0.966	0.642
ISIScox (SCAD)	0.989	0.972	0.986	0.940
Lasso	0.968	0.996	0.967	0.990
Elastic net	0.968	0.993	0.967	0.983

Screening results for $p_{ini} = \lfloor n/\log(n) \rfloor = 37$

k = 30

	m = 0		m = 500	
	1-FPR	TPR	1-FPR	TPR
CUS	0.975	0.427	0.963	0.091
SIScox	0.975	0.436	0.966	0.148
ISIScox (SCAD)	0.978	0.452	0.978	0.213
Lasso	0.977	0.532	0.970	0.269
Elastic net	0.976	0.477	0.966	0.157

Application: Acute myeloid leukaemia

Data: Metzeler et al. (2008), GEO accession GSE12417

- 242 cytogenetically normal AML patients
  - training data: 163 Affymetrix HG-U133 A&B chips
  - test data: 79 Affymetrix HG-U133 plus 2.0 chips
- 44754 probe sets
- Median survival time: 9.6 months in training and 17.6 months in test set, with censoring 37% in training and 40% in test set

Performance assessment

- Model sparseness: number of selected probe sets
- Predictive accuracy: integrated Brier score (IBS, Graf et al. 1999) and associated R^2_{IBS} on independent test data set

Application: Acute myeloid leukaemia

Data: Metzeler et al. (2008), GEO accession GSE12417

- 242 cytogenetically normal AML patients
  - training data: 163 Affymetrix HG-U133 A&B chips
  - test data: 79 Affymetrix HG-U133 plus 2.0 chips
- 44754 probe sets
- Median survival time: 9.6 months in training and 17.6 months in test set, with censoring 37% in training and 40% in test set

Performance assessment

- Model sparseness: number of selected probe sets
- Predictive accuracy: integrated Brier score (IBS, Graf et al. 1999) and associated R^2_{IBS} on independent test data set

Application: Acute myeloid leukaemia

Results: $p_{ini} = n/\log(n) = 31$

	selected probe sets	IBS	R^2_{IBS}
SCAD with initial screening method:			
SIScox	5	0.212	0.060
CUS	3	0.201	0.108
Lasso	4	0.199	0.119
Elastic net	6	0.207	0.083
ISIScox (SCAD)	6	0.206	0.087
Lasso	22	0.195	0.137
Elastic net	22	0.195	0.133
Metzeler et al. (2008)	86	0.211	0.066

Conclusions

- **Pre-Screening is an important** but under-researched aspect of fitting sparse models to high-dimensional data.
- Independence screening methods for linear models are problematic for Cox models (**omitted-covariates problem**)
  - move to parametric survival models
  - for Cox models: ongoing research.
    Idea: Replace the variable of interest by the residuals from a linear regression of it on all other variables (Potter 2005)
- Iterative independence screening **(ISIS) works better** than one-step approach, but is slow and does not solve the omitted-covariates problem.
- In our simulations and application, **lasso performed best** as a screening method.
  - No screening at all - using lasso or elastic net, rather than SCAD - performed even better in terms of prediction accuracy.

Conclusions

- **Pre-Screening is an important** but under-researched aspect of fitting sparse models to high-dimensional data.
- Independence screening methods for linear models are problematic for Cox models (**omitted-covariates problem**)
  - move to parametric survival models
  - for Cox models: ongoing research.
    Idea: Replace the variable of interest by the residuals from a linear regression of it on all other variables (Potter 2005)
- Iterative independence screening **(ISIS) works better** than one-step approach, but is slow and does not solve the omitted-covariates problem.
- In our simulations and application, **lasso performed best** as a screening method.
  - No screening at all - using lasso or elastic net, rather than SCAD - performed even better in terms of prediction accuracy.

Conclusions

- **Pre-Screening is an important** but under-researched aspect of fitting sparse models to high-dimensional data.
- Independence screening methods for linear models are problematic for Cox models (**omitted-covariates problem**)
  - move to parametric survival models
  - for Cox models: ongoing research.
    Idea: Replace the variable of interest by the residuals from a linear regression of it on all other variables (Potter 2005)
- Iterative independence screening **(ISIS) works better** than one-step approach, but is slow and does not solve the omitted-covariates problem.
- In our simulations and application, **lasso performed best** as a screening method.
  - No screening at all - using lasso or elastic net, rather than SCAD - performed even better in terms of prediction accuracy.

Conclusions

- **Pre-Screening is an important** but under-researched aspect of fitting sparse models to high-dimensional data.
- Independence screening methods for linear models are problematic for Cox models (**omitted-covariates problem**)
    - move to parametric survival models
    - for Cox models: ongoing research.
      Idea: Replace the variable of interest by the residuals from a linear regression of it on all other variables (Potter 2005)
- Iterative independence screening **(ISIS) works better** than one-step approach, but is slow and does not solve the omitted-covariates problem.
- In our simulations and application, **lasso performed best** as a screening method.
    - No screening at all - using lasso or elastic net, rather than SCAD - performed even better in terms of prediction accuracy.

References

- **Bühlmann P, Kalisch M, & Maathuis MH (2009)**, Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithm. arXiv:0906.3204v1.
- **Fan J & Lv J (2008)**, Sure independence screening for ultra-high dimensional feature space (with discussion). JRSSB 70:849-911.
- **Fan J & Song R (2009)**, Sure independence screening in generalized linear models with NP-dimensionality. arXiv:0903.5255.
- **Fan J, Samworth R & Wu Y (2009)**, Ultra-dimensional variable selection via independent learning: beyond the linear model. Journal of Machine Learning Research, to appear.
- **Gerds TA & Schumacher M (2001)**, On functional misspecification of covariates in the Cox regression model. Biometrika 88:572-580.
- **Potter DM (2005)**, A permutation test for inference in logistic regression with small- and moderate-sized data sets. Statistics in Medicine 24:693-708.
- **Tibshirani R (2009)**, Univariate shrinkage in the Cox model for high dimensional data. SAGMB 8:21.