Sparse CCA using Lasso

Anastasia Lykou & Joe Whittaker

Department of Mathematics and Statistics,
Lancaster University

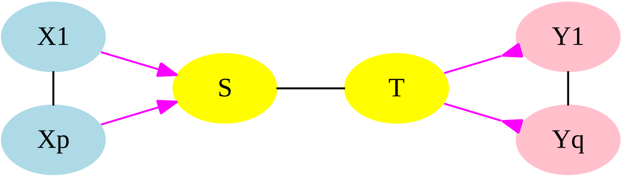
July 23, 2008

Outline

Motivation

- SCCA
  - improve the interpretation of CCA
  - sparse principal component analysis (SCoTLASS by Jolliffe et al. (2003) and SPCA by Zou et al. (2004))
  - interesting data sets (market basket analysis)
- Sparsity
  - shrinkage and model selection simultaneously (may reduce the prediction error, can be extended to high-dimensional data sets)

Canonical Correlation Analysis



- seek linear combinations $S = \boldsymbol{\alpha}^T X$ and $T = \boldsymbol{\beta}^T Y$ such that $\rho = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathrm{corr}(S, T)$
- S, T are the canonical variates
- $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are called conical loadings
- Standard solution through eigen decomposition.

1st dimension

Theorem1

Let α, β be p, q dimensional vectors, respectively.

$$(\widehat{\alpha}, \widehat{\beta}) = argmin_{\alpha,\beta} \left\{ \text{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y}) \right\},$$
$$\text{subject to} \quad \alpha^T \text{var}(\boldsymbol{X})\alpha = \beta^T \text{var}(\boldsymbol{Y})\beta = 1.$$

Then $\widehat{\alpha}, \widehat{\beta}$ are proportional to the first dimensional ordinary canonical loadings.

2nd dimension

Theorem2

Let α, β be p, q dimensional vectors.

$$(\hat{\alpha}, \hat{\beta}) = argmin_{\alpha,\beta} \left\{ \text{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y}) \right\},$$

$$\text{st} \quad \alpha^T \text{var}(\boldsymbol{X})\alpha = \beta^T \text{var}(\boldsymbol{Y})\beta = 1 \;\; \text{and}$$
$$\alpha_1^T \text{var}(\boldsymbol{X})\alpha = \beta_1^T \text{var}(\boldsymbol{Y})\beta = 0$$

where α_1, β_1 are the first canonical loadings.
Then, $\hat{\alpha}, \hat{\beta}$ are proportional to the second dimensional ordinary canonical loadings.

The theorems establish an Alternating Least Squares algorithm for CCA.

2nd dimension

Theorem2

Let α, β be p, q dimensional vectors.

$$(\widehat{\alpha}, \widehat{\beta}) = argmin_{\alpha,\beta} \left\{ \mathrm{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y}) \right\},$$

$$\text{st} \quad \alpha^T \mathrm{var}(\boldsymbol{X})\alpha = \beta^T \mathrm{var}(\boldsymbol{Y})\beta = 1 \quad \text{and}$$

$$\alpha_1^T \mathrm{var}(\boldsymbol{X})\alpha = \beta_1^T \mathrm{var}(\boldsymbol{Y})\beta = 0$$

where α_1, β_1 are the first canonical loadings.
Then, $\widehat{\alpha}, \widehat{\beta}$ are proportional to the second dimensional ordinary canonical loadings.

The theorems establish an Alternating Least Squares algorithm for CCA.

ALS for CCA

Let the objective function be

$$\begin{aligned} Q(\alpha, \beta) &= \mathrm{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y}) \\ \text{subject to} \quad \alpha^T \mathrm{var}(\boldsymbol{X})\alpha &= \beta^T \mathrm{var}(\boldsymbol{Y})\beta = 1. \end{aligned}$$

Q is continuous with closed and bounded domain $\Rightarrow$ Q attains its infimum

ALS algorithm

- Given $\widehat{\alpha}$
- $\widehat{\beta} = \arg\min_\beta Q(\widehat{\alpha}, \beta)$ subject to $\mathrm{var}(\beta^T \boldsymbol{Y}) = 1)$
- Given $\widehat{\beta}$
- $\widehat{\alpha} = \arg\min_\alpha Q(\alpha, \widehat{\beta})$ subject to $\mathrm{var}(\alpha^T \boldsymbol{X}) = 1)$

Q decreases over the iterations and is bounded from below $\Rightarrow$ Q converges.

ALS for CCA

Let the objective function be

$$Q(\alpha,\beta) = \mathrm{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y})$$
$$\text{subject to} \quad \alpha^T \mathrm{var}(\boldsymbol{X})\alpha = \beta^T \mathrm{var}(\boldsymbol{Y})\beta = 1.$$

Q is continuous with closed and bounded domain $\Rightarrow$ Q attains its infimum

ALS algorithm

- Given $\widehat{\alpha}$
- $\widehat{\beta} = \arg\min_{\beta} Q(\widehat{\alpha},\beta)$ subject to $\mathrm{var}(\beta^T \boldsymbol{Y}) = 1)$
- Given $\widehat{\beta}$
- $\widehat{\alpha} = \arg\min_{\alpha} Q(\alpha,\widehat{\beta})$ subject to $\mathrm{var}(\alpha^T \boldsymbol{X}) = 1)$

Q decreases over the iterations and is bounded from below $\Rightarrow$ Q converges.

ALS for CCA

Let the objective function be

$$Q(\alpha, \beta) = \text{var}(\alpha^T \boldsymbol{X} - \beta^T \boldsymbol{Y})$$
$$\text{subject to} \quad \alpha^T \text{var}(\boldsymbol{X})\alpha = \beta^T \text{var}(\boldsymbol{Y})\beta = 1.$$

Q is continuous with closed and bounded domain $\Rightarrow$ Q attains its infimum

ALS algorithm

- Given $\widehat{\alpha}$
- $\widehat{\beta} = \arg\min_{\beta} Q(\widehat{\alpha}, \beta)$ subject to $\text{var}(\beta^T \boldsymbol{Y}) = 1)$
- Given $\widehat{\beta}$
- $\widehat{\alpha} = \arg\min_{\alpha} Q(\alpha, \widehat{\beta})$ subject to $\text{var}(\alpha^T \boldsymbol{X}) = 1)$

Q decreases over the iterations and is bounded from below $\Rightarrow$ Q converges.

Lasso (least absolute shrinkage and selection operator)

- Introduced by Tibshirani (1996)
- Imposes the L_1 norm on the linear regression coefficients.

Lasso

$$\widehat{\beta}_{lasso} = argmin_{\beta} \left\{ var(\mathbf{Y} - \beta^T \boldsymbol{X}) \right\}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

- The L_1 norm properties shrink the coefficients towards zero and exactly to zero if t is small enough.

Lasso algorithms available in the literature

- Lasso by Tibshirani

  Expresses the problem as a least squares problem with 2^p inequality constraints

  Adapts the NNLS algorithm

- Lars-Lasso

  A modified version of Lars algorithm introduced by Efron et al. (2004)

  Lasso estimates are calculated such that the angle between the active covariates and the residuals is always equal.

Proposed algorithm

Lasso with positivity constraints

Suppose that the sign of the coefficients does not change during shrinkage of the coefficients

Positivity Lasso

$$\widehat{\beta}_{lasso} = argmin_{\beta} \left\{ \text{var}(\boldsymbol{Y} - \beta^T \boldsymbol{X}) \right\}$$

$$\text{subject to } s_0^t \beta \leq t \text{ and } s_{0j}\beta_j \geq 0 \text{ for } i = 1 \ldots, p$$

where s_0 is the sign of the OLS estimate.

- simple algorithm, but quite general
- restricted version of Lasso algorithms, since the sign of the coefficients cannot change
- up to $p + 1$ constraints imposed, $<< 2^p$ constraints of Tibshirani's Lasso

Numerical solution

The solution is given through quadratic programming methods,

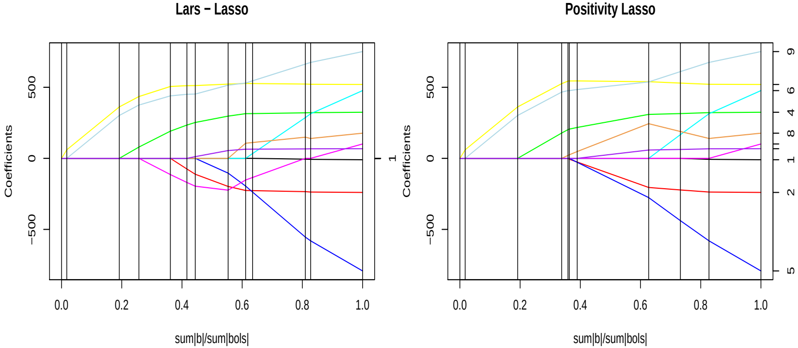
Positivity Lasso solution

$$\widehat{\beta} = b_0 - \lambda \mathsf{var}(X)^{-1} s_0 + \mathsf{var}(X)^{-1} \operatorname{diag}(s_0)\boldsymbol{\mu}$$

- b_0 is the OLS estimate.
- λ is the shrinkage parameter and there is a one to one correspondence between the λ and t
- μ is zero for active and positive for nonactive coefficients
- parameters λ and $\boldsymbol{\mu}$ are calculated satisfying the KKT conditions under the positivity constraints

Diabetes data set

- 442 observations
- age, sex, body mass index, average blood pressure and six blood serum measurements
- disease progression one year after baseline



Simulation studies

We simulate 200 data sets consisting of 100 observations each from the following model,

$$\boldsymbol{Y} = \boldsymbol{\beta}^T \boldsymbol{X} + \sigma\epsilon, \quad \mathrm{corr}(X_i, X_j) = \rho^{|i-j|}$$

Dataset	n	p	$\boldsymbol{\beta}$	σ	ρ
1	100	8	$(3, 1.5, 0, 0, 2, 0, 0, 0)^T$	3	0.50
2	100	8	$(3, 1.5, 0, 0, 2, 0, 0, 0)^T$	3	0.90
3	100	8	$0.85 \forall j$	3	0.50
4	100	8	$(5, 0, 0, 0, 0, 0, 0, 0)^T$	2	0.50

Table: Proportions of the cases the correct model selected.

Dataset	Tibs-Lasso	Lars-Lasso	Pos-Lasso
1	0.06	0.13	0.14
2	0.02	0.04	0.04
3	0.84	0.89	0.87
4	0.09	0.19	0.19

Table: Proportions of agreement between Pos-Lasso and

Dataset	Tibs-Lasso	Lars-Lasso
1	0.76	0.83
2	0.63	0.65
3	0.95	0.98
4	0.77	0.78

ALS for CCA and Lasso

First dimension

Given the canonical variate $T = \beta^T \mathbf{Y}$,

$$\widehat{\alpha} = \arg\min_{\alpha} \left\{ var(T - \alpha^T \boldsymbol{X}) \right\}$$
$$\text{st } var(\alpha^T \boldsymbol{X}) = 1 \text{ and } ||\alpha||_1 \leq t$$

We seek an algorithm solving this optimization problem

or

Modify the Lasso algorithm in order to incorporate the equality constraint.

ALS for CCA and Lasso

First dimension

Given the canonical variate $T = \beta^T \mathbf{Y}$,

$$\widehat{\alpha} = \arg\min_{\alpha} \left\{ var(T - \alpha^T \boldsymbol{X}) \right\}$$
$$\text{st } var(\alpha^T \boldsymbol{X}) = 1 \text{ and } ||\alpha||_1 \leq t$$

We seek an algorithm solving this optimization problem

or

Modify the Lasso algorithm in order to incorporate the equality constraint.

ALS for CCA and Lasso

First dimension

Given the canonical variate $T = \beta^T \mathbf{Y}$,

$$\widehat{\alpha} = \arg\min_{\alpha} \left\{ var(T - \alpha^T \boldsymbol{X}) \right\}$$
$$\text{st } var(\alpha^T \boldsymbol{X}) = 1 \text{ and } ||\alpha||_1 \leq t$$

We seek an algorithm solving this optimization problem

or

Modify the Lasso algorithm in order to incorporate the equality constraint.

ALS for CCA and Lasso

- Tibshirani's Lasso

  NNLS algorithm cannot incorporate the equality constraint
- Lars Lasso

  the equality constraint violates the equiangular condition
- Positivity Lasso

  by additionally imposing positivity constraints the above optimization problem can be solved.

ALS for CCA and Lasso

- Tibshirani's Lasso

  NNLS algorithm cannot incorporate the equality constraint
- Lars Lasso

  the equality constraint violates the equiangular condition
- Positivity Lasso

  by additionally imposing positivity constraints the above optimization problem can be solved.

ALS for CCA and Lasso

- Tibshirani's Lasso
  NNLS algorithm cannot incorporate the equality constraint
- Lars Lasso
  the equality constraint violates the equiangular condition
- Positivity Lasso
  by additionally imposing positivity constraints the above optimization problem can be solved.

SCCA with positivity

First dimension

$$\min_{\boldsymbol{\alpha}} \left\{ \text{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \text{st} \quad \boldsymbol{\alpha}^T \text{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 1,$$

$$\text{and} \quad s_0^T \alpha \leq t, \quad s_{0j} \alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

- The entire Lasso path is derived by considering KKT conditions.
- Cross-validation methods select the shrinkage level applied.
- α_{sp} and β_{sp} for each set of variables are derived alternately until the corr(S_{sp}, T_{sp}) converges to its maximum.

SCCA with positivity

First dimension

$$\min_{\boldsymbol{\alpha}} \left\{ \text{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \text{ st } \quad \boldsymbol{\alpha}^T \text{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 1,$$

$$\text{and} \quad s_0^T \boldsymbol{\alpha} \leq t, \quad s_{0j} \alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

- The entire Lasso path is derived by considering KKT conditions.
- Cross-validation methods select the shrinkage level applied.
- α_{sp} and β_{sp} for each set of variables are derived alternately until the corr(S_{sp}, T_{sp}) converges to its maximum.

SCCA with positivity

First dimension

$$\min_{\boldsymbol{\alpha}} \left\{ \mathrm{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \text{st} \quad \boldsymbol{\alpha}^T \mathrm{var}(\boldsymbol{X})\boldsymbol{\alpha} = 1,$$
$$\text{and} \quad s_0^T \boldsymbol{\alpha} \leq t, \quad s_{0j}\alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

- The entire Lasso path is derived by considering KKT conditions.
- Cross-validation methods select the shrinkage level applied.
- α_{sp} and β_{sp} for each set of variables are derived alternately until the $\mathrm{corr}(S_{sp}, T_{sp})$ converges to its maximum.

SCCA with positivity

Second dimension

$$\min_{\boldsymbol{\alpha}} \left\{ \mathrm{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \quad \text{st} \quad \boldsymbol{\alpha}^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 1, \quad \boldsymbol{\alpha}_1^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 0,$$
$$\text{and} \quad s_0^T \alpha \leq t, \quad s_{0j} \alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

where α_1 is the first dimensional loading.

- Cross-validation methods select the shrinkage level.
- Again alternating algorithm derives the second dimensional canonical loadings

SCCA with positivity

Second dimension

$$\min_{\boldsymbol{\alpha}} \left\{ \mathrm{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \quad \text{st} \quad \boldsymbol{\alpha}^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 1, \quad \alpha_1^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 0,$$
$$\text{and} \quad s_0^T \boldsymbol{\alpha} \leq t, \quad s_{0j} \alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

where α_1 is the first dimensional loading.

- Cross-validation methods select the shrinkage level.
- Again alternating algorithm derives the second dimensional canonical loadings

SCCA with positivity

Second dimension

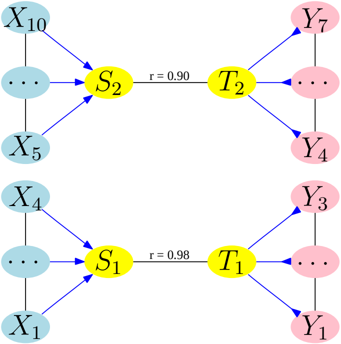
$$\min_{\boldsymbol{\alpha}} \left\{ \mathrm{var}(T - \boldsymbol{\alpha}^T \boldsymbol{X}) \right\} \quad \text{st} \quad \boldsymbol{\alpha}^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 1, \quad \alpha_1^T \mathrm{var}(\boldsymbol{X}) \boldsymbol{\alpha} = 0,$$
$$\text{and} \quad s_0^T \boldsymbol{\alpha} \leq t, \quad s_{0j} \alpha_j \geq 0 \quad \text{for} \quad j = 1, \ldots, p$$

where α_1 is the first dimensional loading.

- Cross-validation methods select the shrinkage level.
- Again alternating algorithm derives the second dimensional canonical loadings

Simulations

We simulate 300 observations of the following model.[1]



Simulations

Variable	1st dim CCA	1st dim SCCA	2nd dim CCA	2nd dim SCCA
$\boldsymbol{X}_1$	0.229	0.248	0.122	0.056
$\boldsymbol{X}_2$	0.350	0.366	-0.052	0
$\boldsymbol{X}_3$	0.337	0.341	0.027	0
$\boldsymbol{X}_4$	0.304	0.298	0.114	0
$\boldsymbol{X}_5$	0.135	0.014	0.198	0.208
$\boldsymbol{X}_6$	-0.037	0	0.381	0.472
$\boldsymbol{X}_7$	-0.052	0	0.212	0.183
$\boldsymbol{X}_8$	-0.052	0	0.205	0.266
$\boldsymbol{X}_9$	-0.111	0	0.166	0.177
$\boldsymbol{X}_{10}$	-0.019	0	0.168	0
$\mathbf{Y}_1$	0.402	0.419	0.112	0.014
$\mathbf{Y}_2$	0.460	0.444	-0.018	0
$\mathbf{Y}_3$	0.309	0.325	0.085	0
$\mathbf{Y}_4$	-0.018	0	0.279	0.421
$\mathbf{Y}_5$	0.032	0	0.183	0.008
$\mathbf{Y}_6$	-0.113	-0.028	0.395	0.361
$\mathbf{Y}_7$	-0.089	-0.025	0.384	0.427
ρ	0.745	0.737	0.654	0.638
RdX (%)	14.2	13.9	13.4	12.6
RdY (%)	16.6	16.4	15	14
Var.Ext of X (%)	25.7	25.5	31.4	30.9
Var.Ext of Y (%)	30	30.1	35	33.8

Summary

Extra work

- Sparse CCA without positivity constraints

  using Lars-Lasso algorithm

Further work

- Compare the performance of SCCA with and without positivity constraints
- Bayesian model selection

  Imposing different Lasso penalties
  Using GVS, Dellaportas et al. (2002)
  Bayesian version of the SCCA

Literature

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12:27–36.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Lawson, C. and Hanson, R. (1974). *Solving Least Square Problems*. Prentice Hall, Englewood Cliffs, NJ.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58:267–288.

Zou, H., Hastie, T., and Tibshirani, T. (2004). Sparse principal component analysis. *to appear, JCGS*.

SCCA

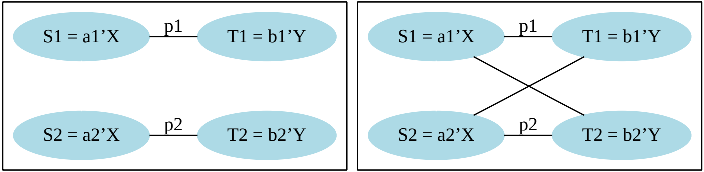


Figure: CCA and SCCA with positivity

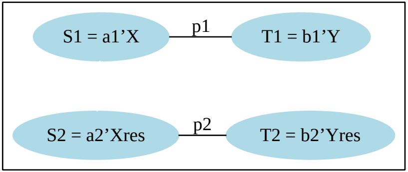


Figure: SCCA without positivity

NNLS

Lawson and Hanson (1974) define the following problems,

LSI problem

LSI problem:	$min_{\beta}\|\|\boldsymbol{Y} - \beta^T \boldsymbol{X}\|\|$ subject to $\mathbf{G}\beta \geq \mathbf{h}$
NNLS problem:	$min_{\beta}\|\|\mathbf{Y} - \beta^T \boldsymbol{X}\|\|$ subject to $\beta \geq 0$
LDP problem:	$min_{\beta}\|\|\beta^T\|\|$ subject to $\mathbf{G}\beta \geq \mathbf{h}$

LSI is equivalent to Lasso $\longrightarrow$ LDP $\longrightarrow$ NNLS