Rank-Based Approach to Optimal Score via Dimension Reduction

Shao-Hsuan Wang
National Taiwan University, Taiwan

Nov 2015

Rank-based measures

- **Kendall's τ**
- **Concordance Index**
- **Rank correlation**

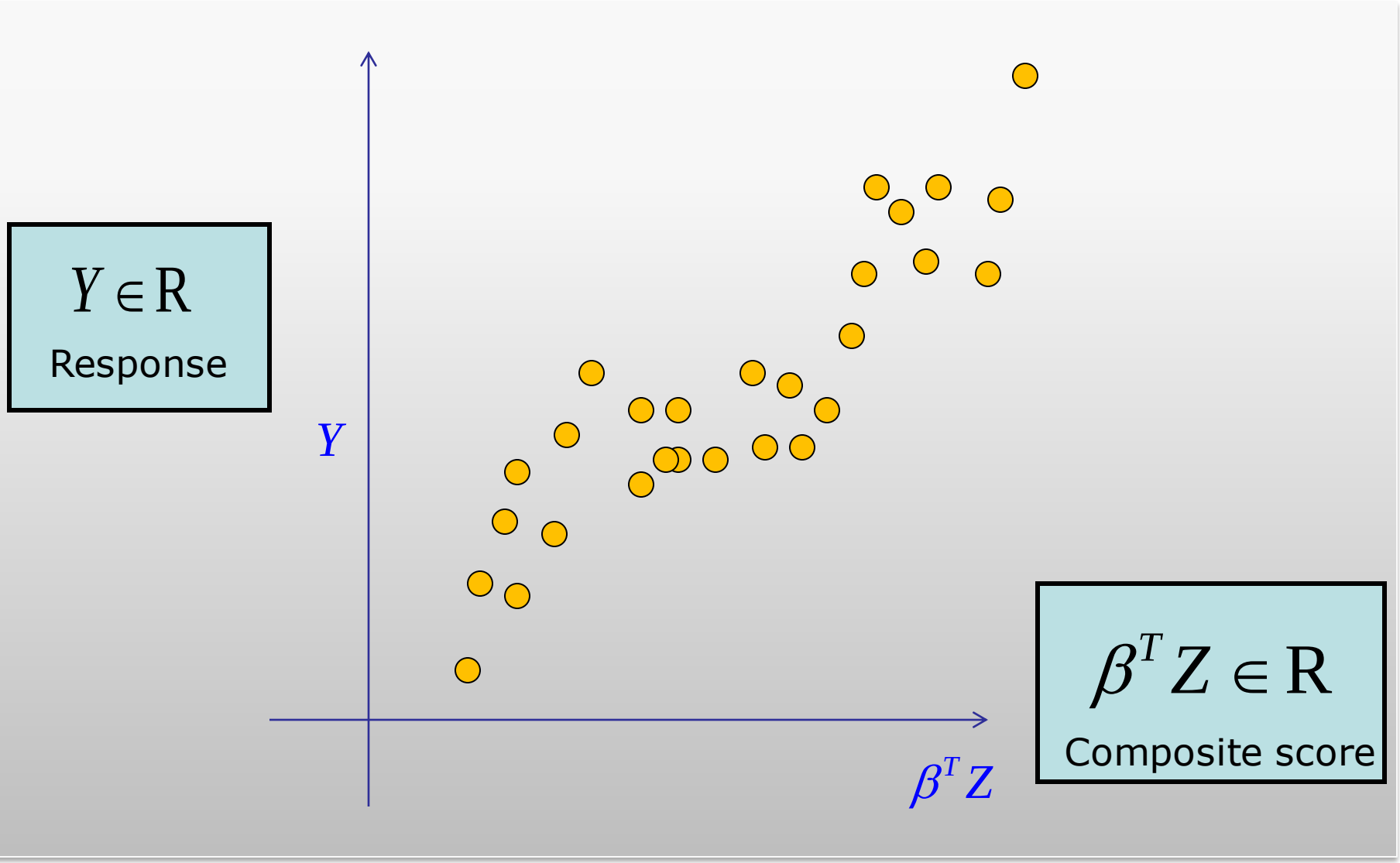
Widely used in medical statistics, epidemiology, economics, and sociology, etc.

Rank-based measures

Regression Model

- Y : **a univariate response**
- $Z = (Z_1, \ldots, Z_p)$: **multiple covariates**

Rank-based measures



Rank-based measures

For pair of observations $(Y_1, \beta^T Z_1)$ **and** $(Y_2, \beta^T Z_2)$ **,**

- **concordant :**

  $Y_1 > Y_2$ and $\beta^T Z_1 > \beta^T Z_2$ $\quad$ $Y_1 < Y_2$ and $\beta^T Z_1 < \beta^T Z_2$

- **discordant :**

  $Y_1 > Y_2$ and $\beta^T Z_1 < \beta^T Z_2$ $\quad$ $Y_1 < Y_2$ and $\beta^T Z_1 > \beta^T Z_2$

Rank-based measures

- **Kendall's τ**

$$\tau = P(Y_1 > Y_2, \beta^T Z_1 > \beta^T Z_2) - P(Y_1 < Y_2, \beta^T Z_1 > \beta^T Z_2)$$

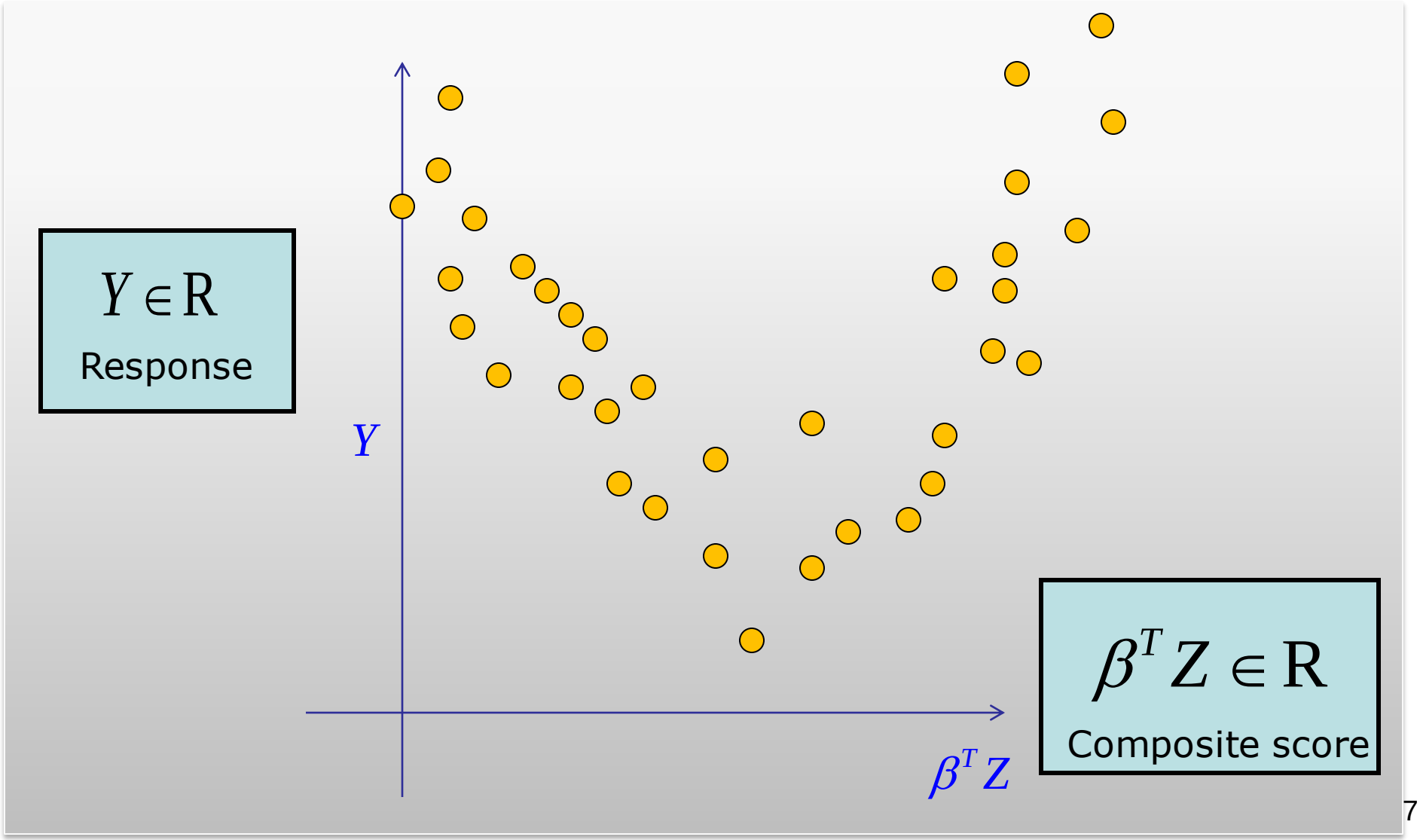
- **Rank correlation**

$$rc = P(Y_1 > Y_2, \beta^T Z_1 > \beta^T Z_2)$$

- **Concordance Index**

$$CI = P(\beta^T Z_1 > \beta^T Z_2 \mid Y_1 > Y_2)$$

Rank-based measures

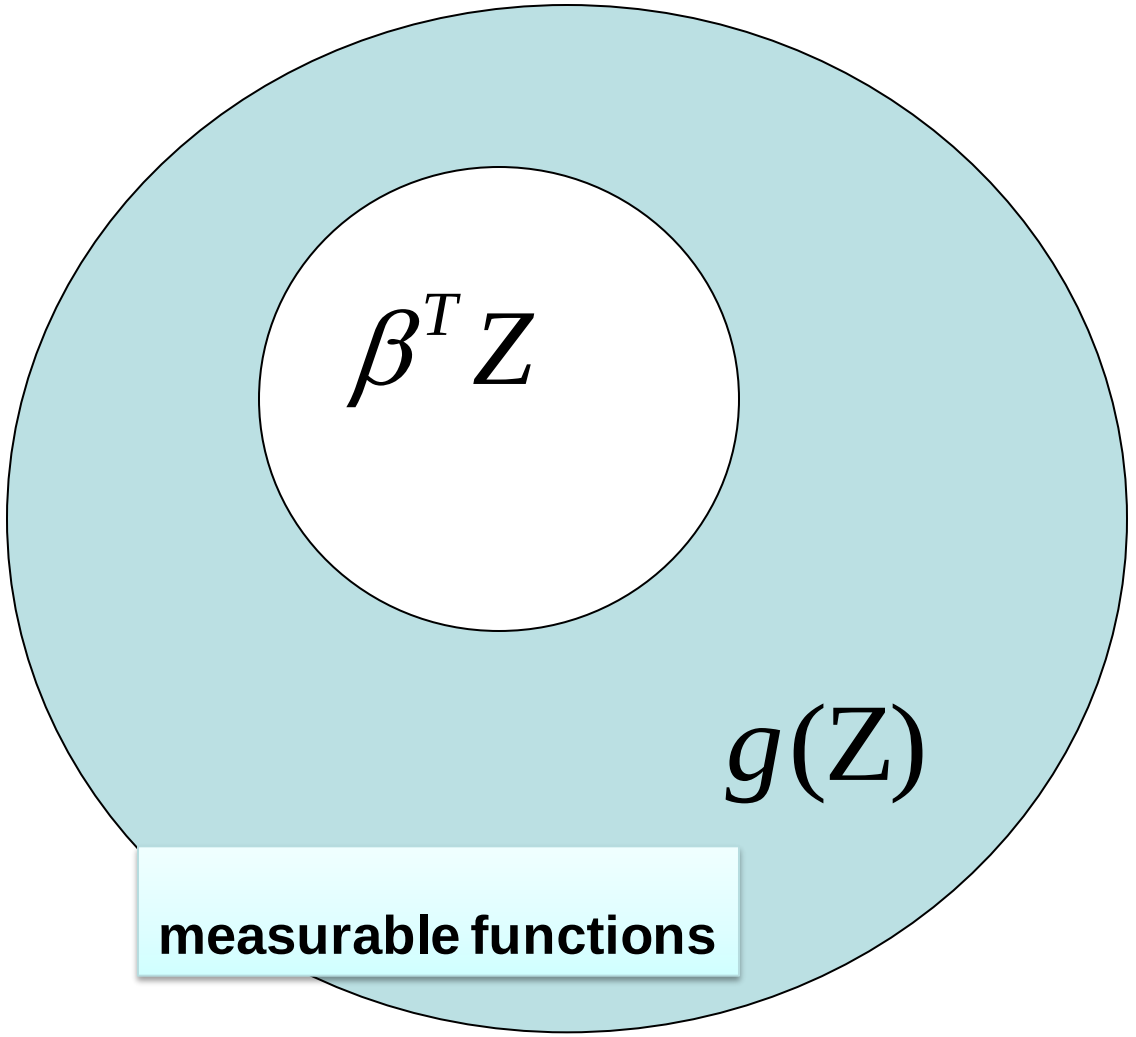


Rank-based measures

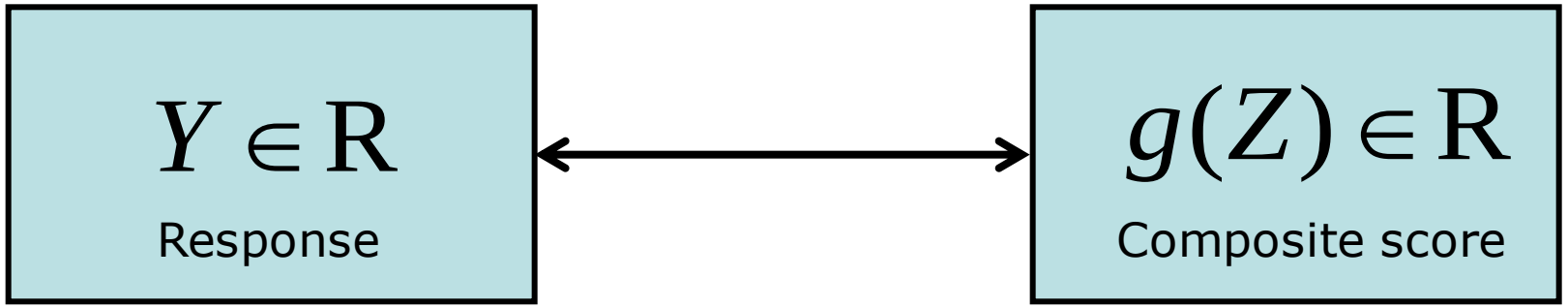
There could not exist a monotonic association !!

Motivation

Composite score



C-max



- **Concordance-index function :**

$$C(g) = P(g(Z_1) > g(Z_2) \mid Y_1 > Y_2)$$

- **C-max :** $C_{\max} = \sup_{g \in F_c} C(g)$

- **Optimal score :** $m(Z)$ such that $m = \sup_{g \in F_c} C(g)$

Intrinsic model
behind Rank-based measures

M1 **Distributional assumption**

：Generalized Regression Model (Han 1987)

M2 **Structural assumption**

: Dimension Reduction (Li 1991, Cook 1991)

Intrinsic model
behind Rank-based measures

M1

a non-degenerate monotonic function on R

$$Y = D \circ G(m_{d_0}(Z), \varepsilon)$$

Intrinsic model
behind Rank-based measures

M1

a non-degenerate monotonic function on R

$$Y = D \circ G(m_{d_0}(Z), \varepsilon)$$

an unspecifed bivariate function strictly increasing at each component for the other one being fixed

Intrinsic model
behind Rank-based measures

M2

$$Y = D \circ G(m_{d_0}(Z), \varepsilon)$$

a multivariate polynomial of the unknown degree d_0

Intrinsic model
behind Rank-based measures

M2

Dimension Reduction

$$m_{d_0}(Z) = m_{d_0 k_0}(B_0^T Z)$$

(1) d_0 **be the smallest degree such that** $Y \perp Z \,|\, m_{d_0}(Z)$

(2) $B_0 = \{\beta_{01}, \ldots, \beta_{0k_0}\}$ **is a basis of the central subspace (CS)**

Model Flexibility

- **Linear regression model** $Y = \beta_0^T Z + \varepsilon$
- **Binary Choice model** $Y = I(\beta_0^T Z + \varepsilon > 0)$
- **Accelerated Failure time model** $\log(Y) = \beta_0^T Z + \varepsilon$
- **Generalized linear regression model (GLM)**
- **Non-monotonic regression model** $Y = (\beta_0^T Z)^2 + \varepsilon$

Types of covariates

- all discrete but continuous covariates
- Covariates which moments could not exist

Theories

Propositions:

(1) Existence $m_{d_0}(Z) \in \arg\max_g C(g)$

(2) Uniqueness $f_{d_0}(Z) \in \arg\max_g C(g) \Leftrightarrow f_{d_0}(Z) = c_1 m_{d_0}(Z) + c_2$

for a ploynomial $f_{d_0}(z)$ of the degree d_0

(3) Optimality $g(Z) \in \arg\max_g C(g) \Leftrightarrow g(Z) = T(m_{d_0}(Z))$

for some monotonic function T

Summary

- $\beta^T Z$ **could not be the best composite score**
- **Model flexibility**
- **Various types of covariates**
- **Optimal score : existence, uniqueness, and optimality**

How to estimate

d_0 : structural degree

k_0 : structural dimension

$S(B_0)$: the central subspace

$m_{d_0 k_0}(B_0 Z)$: the optimal score

$C_{\max}$: the C-max

Estimation Procedure

Estimation Procedure

Step1 Derive $\hat{m}_d(Z)$ by maximizing the concordance index function via the generalized single-index form of the polynomial

Tips:

(1) $$m_d(Z) = \sum_{r=0}^{d} \sum_{r_1+\cdots+r_p=r} c_{r_1\cdots r_p} \prod_{j=1}^{p} Z^{r_j} = \eta^T \underset{\sim}{Z}$$

(2) $$C_n(\mathrm{m}_d(Z)) = C_{0n}(\eta) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(\eta^T \underset{\sim}{\mathrm{Z}}_i > \eta^T \underset{\sim}{\mathrm{Z}}_j, \mathrm{Y}_i > \mathrm{Y}_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I(\mathrm{Y}_i > \mathrm{Y}_j)}$$

Estimation Procedure

Step 2 Apply the outer grandient approach to obtain $\hat{\mathrm{B}}_k$

Tips : (1)

$$m_{d_0}(\mathrm{u}) = m_{d_0 k_0}(B_0^T u)$$

(2)

$$\mathrm{col}(S(B_0)) = \mathrm{col}\Big(\int_{u \in R^p} \nabla m_{d_0}(u)(\nabla m_{d_0}(\mathrm{u}))^T dW(\mathrm{u}) \Big)$$

Estimation Procedure

Step 3 Derive the estimator of $m_{dk}(B_k^T Z)$

Tips : **(1)** $$\hat{Z} = \hat{B}_k^T Z$$

(2) $$\hat{\gamma} = \arg\max_{\eta} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(\eta^T \hat{\underset{\sim}{Z}}_i > \eta^T \hat{\underset{\sim}{Z}}_j, \mathrm{Y}_i > \mathrm{Y}_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I(\mathrm{Y}_i > \mathrm{Y}_j)}$$

(3) $$\hat{m}_{dk}(\hat{\mathrm{B}}_k^T \mathrm{Z}) = \hat{\gamma}^T \hat{Z}$$

Estimation Procedure

Step 4 Adopt the concordance-based generalized BIC to estimate

$d_0, k_0, S(B_0), m_{d_0 k_0}(B_0^T Z),$ and $\mathrm{C}_{\max}$

Tips : (1)

$$IC(d,k) = nC_n(\hat{\mathrm{m}}_{dk}(\hat{\mathrm{B}}_k^T Z)) - \frac{\log n}{2}(\mathrm{C}_k^{k+d} - 1)$$

with $IC(0,k) = 1/2$

(2)

$$(\hat{\mathrm{d}}, \hat{\mathrm{k}}) = \underset{0 \le d, 1 \le k \le p-1}{\arg\max}\ IC(d,k)$$

Asymptotic results

- **Consistent model selection**

--- parsimonious model among the class of Correct models (d_0, k_0)

- $\sqrt{n}$**-consistency of estimators of** $S(B_0)$ **and** $m_{d_0 k_0}(B_0^T Z)$

- **Asymptotic normality of estimators of** $C_{\max}$

Wine data

- **Vinho verde wine : red wine and white wine (from the Minho Region of Northern Portugal)**

- **Collected from May/2004 -February/2007**

- **Red wine : sample size (n)=1599**
  White wine : n=4898

- **Physicochemical and sensory tests**

Wine data

Response (Y):

Preferences 0 (bad) -10 (excellent)

11 Covariates (Z) :

fixed acidity, volatile acidity,

citric acid, residual sugar,

chlorides, free sulfur dioxide,

total sulfur dioxide, density,

PH, sulphates, and alcohol

Wine data

TABLE 1

IC values and the corresponding C-index for two wine datasets

	Red wine			White wine				
(d,k)	C-index	IC	(d,k)	C-index	IC	(d,k)	C-index	IC
(1,1)	**0.800**	**1276.9**	(1,1)	0.762	3730.7	(3,1)	0.754	3680.8
(2,1)	0.717	1140.0	(2,1)	0.753	3682.8	(3,2)	0.756	3664.7
(2,2)	0.763	1202.0	(2,2)	0.765	3727.5	(3,3)	0.770	3692.2
(2,3)	0.783	1219.7	**(2,3)**	**0.790**	**3834.3**	(3,4)	0.774	3647.7
(2,4)	0.789	1211.1	(2,4)	0.793	3826.1	(3,5)	0.782	3600.7
(2,5)	0.799	1204.4	(2,5)	0.793	3794.0	(3,6)	0.786	3499.9
(2,6)	0.827	1223.7	(2,6)	0.796	3785.4	(3,7)	0.792	3374.7
(2,7)	0.832	1202.0	(2,7)	0.796	3752.9	(3,8)	0.794	3196.3
(2,8)	0.833	1170.2	(2,8)	0.798	3722.9	(3,9)	0.805	3014.7
(2,9)	0.837	1140.2	(2,9)	0.801	3696.0	(3,10)	0.807	2743.3
(2,10)	0.837	1099.1	(2,10)	0.804	3663.5	(3,11)	0.807	2402.8
(2,11)	0.838	1056.4	(2,11)	0.804	3611.3			

Wine data

TABLE 2

The estimated basis of the central subspace for Red wine data with bootstrap standard error (s.e.)

	Variables									
	FACID	VACID	CACID	SUGAR	CHLOR	FSDIOX	TSDIOX	DEN	pH	SULPH	ALCOH
	0.16	**-0.45**	**-0.21**	0.06	**-0.14**	**0.11**	**-0.27**	-0.02	-0.08	**0.31**	**0.61**
	(0.108)	(0.062)	(0.075)	(0.064)	(0.064)	(0.054)	(0.077)	(0.100)	(0.071)	(0.058)	(0.062)

TABLE 3

The estimated basis of the central subspace for White wine data with bootstrap standard error (s.e.)

	Variables										
	FACID	VACID	CACID	SUGAR	CHLOR	FSDIOX	TSDIOX	DEN	pH	SULPH	ALCOH
$\hat{\beta}_1$	-0.08	**0.83**	-0.08	-0.20	-0.03	-0.00	-0.02	-0.12	**0.21**	-0.03	-0.13
	(0.07)	(0.164)	(0.109)	(0.109)	(0.090)	(0.062)	(0.078)	(0.074)	(0.084)	(0.083)	(0.073)
$\hat{\beta}_2$	**-0.11**	-0.12	-0.03	**-0.35**	0.04	-0.08	0.07	**0.57**	**-0.23**	-0.06	**-0.14**
	(0.034)	(0.074)	(0.054)	(0.056)	(0.037)	(0.044)	(0.044)	(0.056)	(0.043)	(0.035)	(0.032)
$\hat{\beta}_3$	**0.25**	-0.13	0.05	**-0.16**	**-0.12**	**-0.17**	**-0.25**	**-0.14**	0.07	**0.07**	**0.52**
	(0.039)	(0.073)	(0.047)	(0.042)	(0.028)	(0.042)	(0.041)	(0.032)	(0.040)	(0.027)	(0.047)

Thank You !