Introduction to Envelope Models

Linquan Ma

Department of Statistics
University of Wisconsin-Madison

September 17, 2018

Outline

Regression with multiple responses

- Linear regression model can be written as:

$$\mathbf{y}_{1\times r} = \mathbf{x}_{1\times p}\boldsymbol{\beta}_{p\times r} + \boldsymbol{\epsilon}$$

where the error vector $\boldsymbol{\epsilon} \in \mathbb{R}^r$ is normally distributed with mean $\mathbf{0}$ and unknown parameter $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma} > 0$, the model has a total of $pr + r(r+1)/2$ unknown parameters.

Regression with multiple responses

- Linear regression model can be written as:

$$\mathbf{y}_{1\times r} = \mathbf{x}_{1\times p}\boldsymbol{\beta}_{p\times r} + \boldsymbol{\epsilon}$$

  where the error vector $\boldsymbol{\epsilon} \in \mathbb{R}^r$ is normally distributed with mean $\mathbf{0}$ and unknown parameter $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma} > 0$, the model has a total of $pr + r(r+1)/2$ unknown parameters.

- Suppose we observe dataset $\Big((\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)\Big)$, we can express them in a matrix form:

$$\mathbf{Y}_{n\times r} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times r} + \boldsymbol{\epsilon}$$

Maximum likelihood estimator

- The log-likelihood of linear model:

$$l(\beta, \Sigma|\mathbf{Y}) = -\frac{n}{2}\det(\Sigma) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\beta)^T + C$$

Maximum likelihood estimator

- The log-likelihood of linear model:

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{Y}) = -\frac{n}{2}\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T + C$$

- Setting partial derivative of $\boldsymbol{\beta}$ to $\mathbf{0}$, we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}\boldsymbol{\Sigma}^{-1} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1} = 0$$

Maximum likelihood estimator

- The log-likelihood of linear model:

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{Y}) = -\frac{n}{2}\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\beta)^T + C$$

- Setting partial derivative of $\boldsymbol{\beta}$ to $\mathbf{0}$, we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}\boldsymbol{\Sigma}^{-1} - \mathbf{X}^T\mathbf{X}\beta\boldsymbol{\Sigma}^{-1} = 0$$

- Since $\boldsymbol{\Sigma}$ is positive definite, we can cancel it on both sides. Hence the MLE of β is

$$\hat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}^T\mathbf{X})^{\dagger}\mathbf{X}^T\mathbf{Y}$$

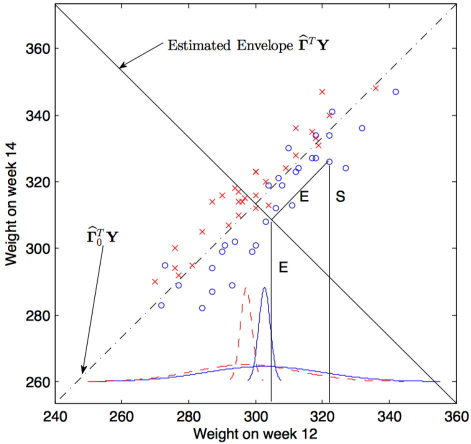
where $\dagger$ indicates the Moore-Penrose inverse.

Remark

The estimation of $\boldsymbol{\beta}$ does not depend on $\boldsymbol{\Sigma}$ at all.

Outline

Motivation



Estimated Envelope $\widehat{\Gamma}^T \mathbf{Y}$
$\widehat{\Gamma}_0^T \mathbf{Y}$
E
S
E
Weight on week 14
Weight on week 12
360
340
320
300
280
260
240
260
280
300
320
340
360

Motivation

- From the graph, we see that some characteristics of the response vector could be unaffected by changes in the predictors.

Motivation

- From the graph, we see that some characteristics of the response vector could be unaffected by changes in the predictors.
- For the ease of notation later, we slightly change the model as

$$\mathbf{y}_{r\times 1} = \boldsymbol{\beta}_{r\times p}\mathbf{x}_{p\times 1} + \boldsymbol{\epsilon}$$

Motivation

- From the graph, we see that some characteristics of the response vector could be unaffected by changes in the predictors.
- For the ease of notation later, we slightly change the model as

$$\mathbf{y}_{r\times 1} = \boldsymbol{\beta}_{r\times p}\mathbf{x}_{p\times 1} + \boldsymbol{\epsilon}$$

- Consider a subspace $\mathcal{E} \subseteq \mathbb{R}^r$ so that ($\mathbf{P}_{\mathcal{E}}$ = projection onto $\mathcal{E}$, $\mathbf{Q}_{\mathcal{E}} = \mathbf{I} - \mathbf{P}_{\mathcal{E}}$)

$$\mathbf{Q}_{\mathcal{E}}\mathbf{Y}|(\mathbf{X}=\mathbf{x}_1) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y}|(\mathbf{X}=\mathbf{x}_2), \forall(\mathbf{x}_1,\mathbf{x}_2) \iff \mathrm{span}(\boldsymbol{\beta}) \subset \mathcal{E}$$

$$\mathbf{P}_{\mathcal{E}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{Y}|\mathbf{X} \iff \boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$$

Motivation

- From the graph, we see that some characteristics of the response vector could be unaffected by changes in the predictors.
- For the ease of notation later, we slightly change the model as

$$\mathbf{y}_{r\times 1} = \boldsymbol{\beta}_{r\times p}\mathbf{x}_{p\times 1} + \boldsymbol{\epsilon}$$

- Consider a subspace $\mathcal{E} \subseteq \mathbb{R}^r$ so that ($\mathbf{P}_{\mathcal{E}}$ = projection onto $\mathcal{E}$, $\mathbf{Q}_{\mathcal{E}} = \mathbf{I} - \mathbf{P}_{\mathcal{E}}$)

$$\mathbf{Q}_{\mathcal{E}}\mathbf{Y}|(\mathbf{X} = \mathbf{x}_1) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y}|(\mathbf{X} = \mathbf{x}_2), \forall(\mathbf{x}_1, \mathbf{x}_2) \iff \mathrm{span}(\boldsymbol{\beta}) \subset \mathcal{E}$$

$$\mathbf{P}_{\mathcal{E}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{Y}|\mathbf{X} \iff \boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$$

- This implies the impact of $\mathbf{X}$ on $\mathbf{Y}$ is concentrated only in $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$. We refer to $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ and $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ informally as material and immaterial part of $\mathbf{Y}$.

Outline

Invariant and reducing subspaces

- A subspace $\mathcal{R}$ of $\mathbb{R}^r$ is an invariant subspace of $\mathbf{M} \in \mathbb{R}^{r \times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$.
- $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, if, in addition, $\mathbf{M}\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$.

Invariant and reducing subspaces

- A subspace $\mathcal{R}$ of $\mathbb{R}^r$ is an invariant subspace of $\mathbf{M} \in \mathbb{R}^{r \times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$.
- $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, if, in addition, $\mathbf{M}\mathcal{R}^\perp \subseteq \mathcal{R}^\perp$.

Proposition 1.

$\mathcal{R}$ reduces $\mathbf{M} \in \mathbb{R}^{r \times r}$ if and only if $\mathbf{M}$ can be written in the form:

$$\mathbf{M} = \mathbf{P}_\mathcal{R}\mathbf{M}\mathbf{P}_\mathcal{R} + \mathbf{Q}_\mathcal{R}\mathbf{M}\mathbf{Q}_\mathcal{R}$$

Invariant and reducing subspaces

- A subspace $\mathcal{R}$ of $\mathbb{R}^r$ is an invariant subspace of $\mathbf{M} \in \mathbb{R}^{r\times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$.
- $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, if, in addition, $\mathbf{M}\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$.

Proposition 1.

$\mathcal{R}$ reduces $\mathbf{M} \in \mathbb{R}^{r\times r}$ if and only if $\mathbf{M}$ can be written in the form:

$$\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$$

Proposition 2.

Assume that $\mathbf{M}_{r\times r}$ is symmetric and has $q \leq r$ distinct eigenvalues, and let $\mathbf{P}_i, i = 1, ..., q$, indicate the projection onto the corresponding eigenspaces. Then $\mathcal{R}$ reduces $\mathbf{M}$ if and only if $\mathcal{R} = \oplus_{i=1}^{q} \mathbf{P}_i\mathcal{R}$

M-envelopes

Definition 1. (Cook et al. (2010))

Let $\mathbf{M}$ be a symmetric matrix, and let $\mathcal{S} \subseteq \mathrm{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{S}$, to be written as $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contains $\mathcal{S}$.

M-envelopes

Definition 1. (Cook et al. (2010))

Let $\mathbf{M}$ be a symmetric matrix, and let $\mathcal{S} \subseteq \mathrm{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{S}$, to be written as $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contains $\mathcal{S}$.

Proposition 3.

$$\mathcal{E}_{\mathbf{M}}(\mathcal{S}) = \oplus_{i=1}^{q} \mathbf{P}_i \mathcal{S}$$

where $\mathbf{P}_i$ is the projection matrix to the i-th eigenspace of $\mathbf{M}$.

Outline

- Now we want to refine the regression model by using an envelope to connect $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.
- Let the columns of the semi-orthogonal matrices $\boldsymbol{\Gamma} \in \mathbb{R}^{u \times r}$ be the base of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{(r-u) \times r}$ be its orthogonal complements. Then, there exist an $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$ such that $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ where $\boldsymbol{\eta}$ contains the coordinates of $\boldsymbol{\beta}$ relative to $\boldsymbol{\Gamma}$.

Envelope model:

$$\mathbf{Y} = \boldsymbol{\Gamma}\eta\mathbf{X} + \epsilon, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$$

- Estimation of the parameters can be carried out by maximum likelihood with u determined by AIC, BIC or other methods.

Maximum likelihood estimation

- The estimated envelope $\hat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ can be represented as

$$\hat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \arg\min_{\boldsymbol{\delta}}(\log|\mathbf{P}_\delta \mathbf{S}_{\mathbf{Y}|\mathbf{X}} \mathbf{P}_\delta|_0 + \log|\mathbf{Q}_\delta \mathbf{S}_{\mathbf{Y}} \mathbf{Q}_\delta|_0)$$

where $|\cdot|_0$ means the product of the non-zero eigenvalues and $\boldsymbol{\delta}$ is a u-dim subspace of $\mathbb{R}^r$. $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{Y}}$ are the sample version of $\mathbf{\Sigma}$ and $\mathrm{var}(\mathbf{Y})$.

Maximum likelihood estimation

- The estimated envelope $\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be represented as

$$\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \arg\min_{\boldsymbol{\delta}}(\log|\mathbf{P}_{\delta}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\delta}|_0 + \log|\mathbf{Q}_{\delta}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\delta}|_0)$$

  where $|\cdot|_0$ means the product of the non-zero eigenvalues and $\boldsymbol{\delta}$ is a u-dim subspace of $\mathbb{R}^r$. $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{Y}}$ are the sample version of $\boldsymbol{\Sigma}$ and $\mathrm{var}(\mathbf{Y})$.
- $\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be estimated through a 1-D algorithm proposed by Cook and Zhang (2016).

Maximum likelihood estimation

- The estimated envelope $\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be represented as

$$\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \arg\min_{\delta}(\log|\mathbf{P}_\delta \mathbf{S}_{\mathbf{Y}|\mathbf{X}} \mathbf{P}_\delta|_0 + \log|\mathbf{Q}_\delta \mathbf{S}_{\mathbf{Y}} \mathbf{Q}_\delta|_0)$$

  where $|\cdot|_0$ means the product of the non-zero eigenvalues and δ is a u-dim subspace of $\mathbb{R}^r$. $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{Y}}$ are the sample version of $\boldsymbol{\Sigma}$ and $\mathrm{var}(\mathbf{Y})$.
- $\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be estimated through a 1-D algorithm proposed by Cook and Zhang (2016).
- Let $\hat{\boldsymbol{\Gamma}}$ be a basis for $\hat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$, estimators of other parameters are:
  - $\hat{\beta}_{env} = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}} \hat{\beta}_{ols}$, which is $\sqrt{n}$ consistent and asymptotically normal.
  - $\hat{\boldsymbol{\Sigma}} = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}} \mathbf{S}_{\mathbf{Y}|\mathbf{X}} \mathbf{P}_{\hat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\hat{\boldsymbol{\Gamma}}} \mathbf{S}_{\mathbf{Y}|\mathbf{X}} \mathbf{Q}_{\hat{\boldsymbol{\Gamma}}}$

Efficiency gain

Proposition 4. (Cook et al. (2010))

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}[\hat{\beta}_{env}]) \leq \mathrm{avar}(\sqrt{n}\mathrm{vec}[\hat{\beta}_{std}])$$

where $\mathrm{avar}(\cdot)$ stands for the asymptotic covariance and $\mathrm{vec}(\cdot)$ stands for the vectorization of a matrix.

Efficiency gain

Proposition 4. (Cook et al. (2010))

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}[\hat{\beta}_{env}]) \leq \mathrm{avar}(\sqrt{n}\mathrm{vec}[\hat{\beta}_{std}])$$

where $\mathrm{avar}(\cdot)$ stands for the asymptotic covariance and $\mathrm{vec}(\cdot)$ stands for the vectorization of a matrix.

Remarks

- Envelope methods are never worse than standard methods.
- Envelope methods will provide the most gain in efficiency when $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ can be constructed from eigenspaces of $\mathbf{\Sigma}$ with relatively small eigenvalues.

Outline

Example

Illustration of an example using R:

```
require(envlp)
set.seed(0411)
num = 200
env_dim <- 5
p = 5
q = 20
sq_err_env <- NULL
sq_err_std <- NULL
for (i in 1:10) {
  GAMMA <- matrix(runif(env_dim * q), nrow = q)
  beta0 <- matrix(runif(p * q, -10, 10), nrow = p)
  beta <- beta0 %*% P(GAMMA)
  Omega <- 0.1 * diag(nrow(GAMMA))
  Omega0 <- 1000 * diag(nrow(GAMMA))
  Sigma_y <- P(GAMMA) %*% Omega %*% P(GAMMA) + Q(GAMMA) %*% Omega0 %*% Q(GAMMA)
  A <- matrix(runif(p ^ 2, -10, 10), nrow = p)
  mu_x <- runif(p, -10, 10)
  Sigma_x <- A %*% t(A)
  X <- mvrnorm(num, mu_x, Sigma_x)
  Y <- X %*% beta + mvrnorm(num, rep(0, q), Sigma_y)
```

Example

```
  u = u.env(X, Y)$u.bic
  env_beta <- t(env(X, Y, u)$beta)
  std_beta <- solve(crossprod(X)) %*% crossprod(X, Y)
  sum((std_beta - beta)^2)
  sq_err_env <- c(sum((env_beta - beta)^2), sq_err_env)
  sq_err_std <- c(sum((std_beta - beta)^2), sq_err_std)
}
mean(sq_err_env)
mean(sq_err_std)
```

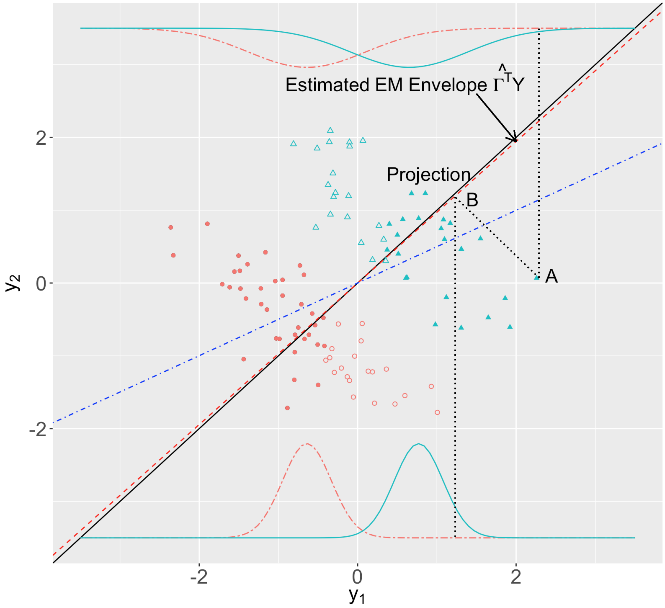
Example

```
  u = u.env(X, Y)$u.bic
  env_beta <- t(env(X, Y, u)$beta)
  std_beta <- solve(crossprod(X)) %*% crossprod(X, Y)
  sum((std_beta - beta)^2)
  sq_err_env <- c(sum((env_beta - beta)^2), sq_err_env)
  sq_err_std <- c(sum((std_beta - beta)^2), sq_err_std)
}
mean(sq_err_env)
mean(sq_err_std)
```

- Output: 0.007500058 14.07488

Outline

Envelope methods with ignorable missing data



Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960.

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300.