STA5001: High Dimensional Statistics

Gong Wenwu 12031299

June 3, 2021

1. Sol. of LASSO problem

$$LASSO: \ \arg\min_{\beta} \frac{1}{2n} \| Y - X\beta \|_2{}^2 + \| \beta \|_1 = \arg\min_{\beta} f_\lambda(\beta), \beta \in \Re^p. \tag{1}$$

- (a). We can induct the LASSO sol. by using subgradient.
  Proof: Note that the gradient $G(\beta) = -X^T(Y - X\beta)/n$ and subgradient of $\| \beta \|_1$ is

$$S(\beta_j) = \begin{cases} sign(\beta_j), & \beta_j \neq 0 \\ [-1, 1], & \beta_j = 0 \end{cases}$$

  " $\Rightarrow$ " Differentiating $f_\lambda(\beta)$ w.r.t. β_j and setting the equation to zero, we obtain $G_j(\beta) + \lambda S(\beta_j) = 0$, i.e.,

  - if $\hat{\beta}_j \neq 0$, $G_j(\hat{\beta}) = -\lambda sign(\hat{\beta}_j)$.
  - if $\hat{\beta}_j = 0$, $G_j(\hat{\beta}) = -\lambda c \leq |\lambda|$, where $|c| \leq 1$.

  " $\Leftarrow$ " If we have $G_j(\beta) + \lambda S(\hat{\beta}_j) = 0$ for any $\hat{\beta}_j \in \beta$ holds, then we have p equations and can be denoted as $G(\beta) + \lambda S(\beta) = 0$, which is equivalent to $\arg\min_\beta f_\lambda(\beta)$. □

- (b). Suppose we have two sols. $\hat{\beta}^1 \neq \hat{\beta}^2$ and $G_j(\hat{\beta}^1) < \lambda, G_j(\hat{\beta}^2) = -\lambda sign(\hat{\beta}^2_j)$ for some j holds. Let $\hat{\beta}^\gamma = (1-\gamma)\hat{\beta}^1 + \gamma\hat{\beta}^2$, then $\hat{\beta}^\gamma$ also a sol. of LASSO problem. However, for fixed j, $G_j(\hat{\beta}^\gamma) + S(\hat{\beta}^\gamma_j) = 0 \Leftrightarrow \hat{\beta}^2_j = 0$, which is contradicted with $G_j(\hat{\beta}^2) = -\lambda sign(\hat{\beta}^2_j)$, so $G_j(\hat{\beta}^2) < \lambda$ and $\hat{\beta}^2_j = 0$ holds, i.e., $\hat{\beta}_j = 0$ for all solutions if there is sol. $\hat{\beta}$ making that $G_j(\hat{\beta}) < \lambda$. □

2. Properties of penalty function

$$\arg\min_{\theta} g(\theta|z,\lambda) = \arg\min_{\theta} \frac{1}{2}(z-\theta)^2 + P_\lambda(|\theta|) \Rightarrow \theta + P_\lambda^{'}(|\theta|) = z \tag{2}$$

Let $f(\theta|\lambda) = \theta + P_\lambda^{'}(|\theta|)$, then we need to compare the intersection point of function $f(\theta|\lambda)$ and z $(constant)$. Since $\hat{\theta}(z) = -\hat{\theta}(-z)$ (Later, I will give a proof) and $\hat{\theta}(z) \geq 0$ when $z \geq 0$ holds (it is obvious if there is sol. for $z \geq 0$). So we only need to calculate $\hat{\theta}(z)$ when $z \geq 0$, i.e., discussing the properties of penalty function when $\theta \geq 0$.

- (a.) If $t_0 = \min_{\theta \leq 0} f(\theta|\lambda) > 0$, we say there is no intersection point when $z < t_0$, so $\hat{\theta}(z|\lambda) = 0$, the penalty function has the sparsity.
- (b.) If $P_\lambda^{'}(|\theta|) = 0$ for $|\theta| \geq 0$, then $f(\theta|\lambda) = \theta$. So the sols. existed assure that $\hat{\theta}(z|\lambda) = z$ when $|\theta| \geq 0$ holds, the penalty function has the approximate unbiasedness.
- (c.) Firstly, we can prove that $\hat{\theta}(z)$ is a odd function, i.e., $\hat{\theta}(z) = -\hat{\theta}(-z)$. Since $(-z-(-\theta))^2 = (\theta - z)^2 = (z-\theta)^2$ and $P_\lambda(|\theta|) = P_\lambda(|-\theta|)$, so, if Eq. 2 holds for $\hat{\theta}(z)$, then $-\hat{\theta}(-z)$ also holds, we say that $\hat{\theta}(z) = -\hat{\theta}(-z)$. Then the odd function and $\hat{\theta}(z) \geq 0$ when $z \geq 0$ holds, assuring that $\arg\min_{\theta \geq 0} f(\theta|\lambda) = 0$, i.e., the penalty function has the continuity.

3. Concentration inequality for the median-of-means estimator

Consider a scenario where we observe $X1, \cdots, Xn \sim F$, the goal is to estimate the mean of the underlying distribution $\mu = E(X) = \int x d_F$. The MoM estimator works as follows: Assume that the sample size$n = k * m$, where k is the number of sub-samples and m is the size of each sub-sample. We first randomly split the data into k sub-sample and compute the mean using each sub-sample, which leads to estimators $\hat{\mu_k}$ and each estimator is based on m observations. Then the MoM estimator $\hat{\mu}_{MoM}$ is defined as the median of all these estimator, i.e., $\hat{\mu}_{MoM} = median\{\hat{\mu_1}, \cdots, \hat{\mu_k}\}$. We divide the MoM problem into three simple steps:

- (a.) Firstly, WTS:

$$Pr(|\hat{\mu}_m - \mu| \geq \frac{2\sigma}{\sqrt{m}}) \leq \frac{1}{4},\ m\ is\ the\ size\ of\ sample.$$

Proof: By Chebyshev's inequality,

$$Pr(|\hat{\mu}_m - \mu| \geq \frac{2\sigma}{\sqrt{m}}) \leq \frac{E(\hat{\mu}_m - \mu)^2}{(2\sigma/\sqrt{m})^2} = \frac{1}{4}.$$

. □

- (b.) Secondly, WTS:

$$Pr(\sum_{j=1}^{k} B_j \geq k/2) \leq (4p(1-p))^{\frac{k}{2}},\ if\ E(B_j) = p < \frac{1}{2}.$$

Proof: Let $\sum_{j=1}^{k} B_j = S_k$, we have $Pr(S_k \geq k/2) = Pr(\exp(tS_k) \geq \exp(tk/2))$ for any t>0 holds. Then by Markov's inequality, we have $Pr(\exp(tS_k) \geq \exp(tk/2)) \leq \exp(-tk/2)E(\exp(tS_k))$. For Bernoulli R.V. B_j, the moment generative function $E(\exp(tS_k)) = (p\exp(t) + (1-p))^k$, then $Pr(S_k \geq k/2) \leq (p\exp(t/2) + (1-p)\exp(-t/2))^k$ for any $t > 0$ holds. Also, we can prove that $g(t) = p\exp(t/2) + (1-p)\exp(-t/2)$ is minimal attainable when $t = \sqrt{\frac{1-p}{p}} (> 1,\ if\ p > \frac{1}{2})$ and the minimal value is $2\sqrt{p(1-p)}$. So, $Pr(\sum_{j=1}^{k} B_j \geq k/2) \leq (4p(1-p))^{\frac{k}{2}}$. □

- (c.) ***Proof:*** Since $\hat{\mu}_{MoM} = median\{\hat{\mu_1}, \cdots, \hat{\mu_k}\}$, then event $\{|\hat{\mu}_{MoM} - \mu| \geq \epsilon\} \subset \{\sum_{j=1}^{k} 1_{(|\hat{\mu}_j - \mu| \geq \epsilon)} \geq k/2\}$, let $B_j = 1_{(|\hat{\mu}_j - \mu| \geq \epsilon)}$, we have $Pr(|\hat{\mu}_{MoM} - \mu| \geq \epsilon) \leq Pr(\sum_{j=1}^{k} B_j \geq k/2)$ holds. We can prove the following inequality by denoting $\epsilon = 2\sigma/\sqrt{m}$ and the results given by (a.), (b.).

$$Pr(|\hat{\mu}_{MoM} - \mu| \geq \frac{2\sigma}{\sqrt{m}}) \leq (\frac{\sqrt{3}}{2})^k. \tag{3}$$

From (a.), $p = E(B_j) = Pr(|\hat{\mu}_j - \mu| \geq \frac{2\sigma}{\sqrt{m}}) \leq \frac{1}{4}$., and (b.) assure that $Pr(\sum_{j=1}^{k} B_j \geq k/2) \leq (4p(1-p))^{\frac{k}{2}} \leq (\frac{\sqrt{3}}{2})^k$, so $Pr(|\hat{\mu}_{MoM} - \mu| \geq \frac{2\sigma}{\sqrt{m}}) \leq (\frac{\sqrt{3}}{2})^k$ holds. □

4. Majorization-minimization algorithm: gradient descent method

Proof:

- (a.) Majorization Step: We can make *Taylor Expansion* of convex function $f(\mathbf{x})$ around point $\mathbf{x}_{i-1}$:

$$f(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f^{'}(\mathbf{x}_{i-1})^{T}(\mathbf{x} - \mathbf{x}_{i-1}) + f^{''}(\mathbf{x}_{i-1})/2 \parallel \mathbf{x} - \mathbf{x}_{i-1} \parallel^2 + Remainder.$$

  Since $f^{''}(\mathbf{x}) \preceq L1_p,\ L \geq \frac{1}{\delta}$, we can say that $g(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{2} \parallel \mathbf{x} - \mathbf{x}_{i-1} \parallel^2 (L1_p - f^{''}(\mathbf{x})) \geq 0$ for all $\mathbf{x}$ and $f(\mathbf{x}_{i-1}) = g(\mathbf{x}_{i-1})$. □

- (b.) Minimization Step: We can differentiate $g(\mathbf{x}_i)$ and set $g^{'}(\mathbf{x}_i) = 0$ to find out the minimizer.

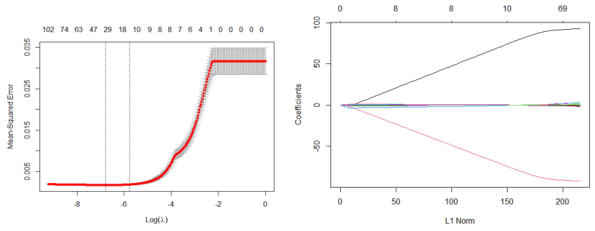
$$g^{'}(\mathbf{x}_i) = f^{'}(\mathbf{x}_{i-1}) + \frac{1}{\delta}(\mathbf{x}_i - \mathbf{x}_{i-1}) = 0 \Rightarrow \mathbf{x}_i = \mathbf{x}_{i-1} - \delta f^{'}(\mathbf{x}_{i-1})$$

  which shows the sol. is a iteration process if $f^{'}(\mathbf{x})$ existed. By the Majorization and Minimization Steps, we assure that it is a MM-algorithm. □

- (c.) Suppose that $\mathbf{x}_{i-1} < \mathbf{x}^* < \mathbf{x}_i$, then $f^{'}(\mathbf{x}_{i-1}) < 0 < f^{'}(\mathbf{x}_i)$ holds. We can expand $f(x)$ around $\mathbf{x}_{i-1}, \mathbf{x}_i$, for fixed $\mathbf{x}^*$, (a.) has shown that $\mathbf{x}_{i-1} - f(\mathbf{x}^*) \leq \frac{1}{2\delta} \parallel \mathbf{x}^* - \mathbf{x}_{i-1} \parallel^2$ and $\mathbf{x}_i - f(\mathbf{x}^*) \leq \frac{1}{2\delta} \parallel \mathbf{x}^i - \mathbf{x}_* \parallel^2$. Then $\mathbf{x}_{i-1} - f(\mathbf{x}^i) \leq \frac{1}{2\delta}\{\parallel \mathbf{x}^* - \mathbf{x}_i \parallel^2 + \parallel \mathbf{x}^{i-1} - \mathbf{x}_* \parallel^2\}$ holds. On the other hands, the convex function assure that $f(\mathbf{x}^i) \leq f(\mathbf{x}^*) + f(\mathbf{x}^{i-1}) - f(\mathbf{x}^i)$, so we have proved that $f(\mathbf{x}^i) \leq f(\mathbf{x}^*) + \frac{1}{2\delta}\{\parallel \mathbf{x}^* - \mathbf{x}_i \parallel^2 + \parallel \mathbf{x}^{i-1} - \mathbf{x}_* \parallel^2\}$. □

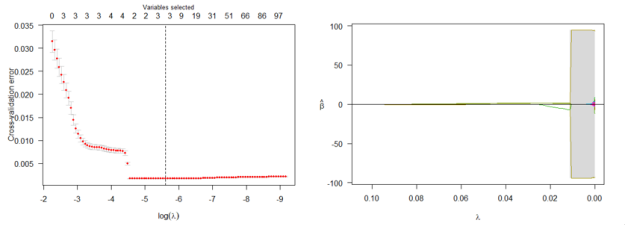
5. Macroeconomic time series

- (a.) The features' name with missing entries are *ACOGNO, ANDENOx, TWEXMMTH,UMCSENTx,VXOCLSx* and we have 122 predictors (UNRATE as response).
- (b.)
  - ***glmnet***: The in-sample $R^2 = 0.9489144$; Two most important macroeconomic variables: *HWIURATIO(88.0127977287329), CLF16OV(-87.6513585414659)* , because the absolute value of these coefficients are bigger than others.

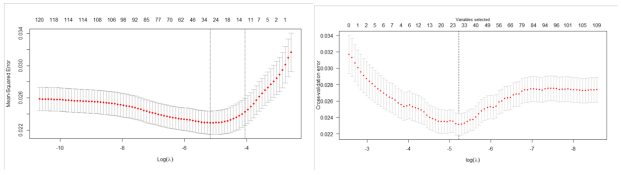


```
> model[["beta"]]@x
  [1]  -0.181705433  -0.210382252  -0.086719606  -0.064433381   0.105847114  -0.009790118
  [7]   0.013053321  -0.042966612  88.012797729 -87.651358541   0.074058970   0.053413149
 [13]   0.093352092   0.013762065   0.001952721  -0.018735783   0.560897313  -0.300923124
 [19]   0.062524125  -0.008939590   0.044532600   0.098415049   0.039843142  -0.002376630
 [25]  -0.230281444  -0.333030697  -0.649342475  -0.097786326   0.278479520  -0.024989192
 [31]   0.048942762   0.119377164
> model[["beta"]]@i
 [1]   0   2   4  13  14  16  17  20  21  22  24  25  26  27  29  33  66  67  69  79  92
[22]  93  98 100 102 108 112 113 114 118 119 121
```

  - ***SCAD***: The in-sample $R^2 = 0.95$; Two most important macroeconomic variables: *UEMPLT5(0.198067904823108), BAAFFM(0.0705408044680082).*



- (c.)
  - ***glmnet***: The in-sample $R^2 = 0.3520272$; *NDMANEMP(-10.360460516), MANEMP(-6.579253670)* .



LASSO SCAD

  - ***SCAD***: The in-sample $R^2 = 0.27$; *COMPAPFFx(0.1294552528), TB3SMFFM(-0.0883961041).*
- (d.) The out-of-sample R^2 of LASSO is 0.187032279566027, and SCAD is 0.26121426174681.

6. Group LASSO

The group-Lasso was proposed to solve the variable selection in the additive model, the penalty function is given by $P_\lambda(\|\theta_j\|_{W_j})$, where $\|\theta_j\|_{W_j} = \sqrt{({\theta_j}^T W_j \theta_j)}$ and we suppose that $W_j = 1_{pj}$.

$$Group\ LASSO\ \arg\min_\theta \frac{1}{2}\| Y - \sum_{j=1}^{J} Z_j\theta_j \|_2^{\ 2} + \lambda \sum_{j=1}^{J} \| \theta_j \|_2 = \arg\min_\theta f(\theta) \Rightarrow \hat{\theta} = ? \tag{4}$$

- (a.) If $\theta_j \neq 0$, the derivative of $\| \theta_j \|_2 = \frac{\theta_j}{\sqrt{{\theta_j}^T\theta_j}} = \frac{\theta_j}{\|\theta_j\|_2}$, then $\arg\min_\theta f(\theta_j)$ equals to $-{Z_j}^T(Y - \sum_{j=1}^{J} Z_j\theta_j) + \lambda\frac{\theta_j}{\|\theta_j\|_2} = 0$. Under the sub-differential form S_j and fixed $\hat{\theta}_j$, we have $\hat{S}_j = \frac{\hat{\theta}_j}{\|\hat{\theta}_j\|_2}$. If $\hat{\theta}_j = 0$, we have $\langle Z_j, Y\rangle < \lambda$, then $\| \hat{S}_j \|_2 = \| \frac{\langle Z_j, Y\rangle}{\lambda} \|_2 \leq 1$. □

- (b.) ***Block CCD Algorithm***: Given the initial parameter value, we introduce $\gamma_j = Y - \sum_{k\neq j} Z_k\theta_k$ and then update the estimate one group at a time(In this case, we don't assume that group-wise orthogonal). From (a.), we have $\arg\min_\theta f(\theta_j) \Rightarrow -Z_j(\gamma_j - Z_j\theta_j) + \lambda S_j = 0$. If $\hat{\theta}_j = 0$, $\| \hat{S}_j \|_2 \leq 1, i.e., \langle Z_j Y\rangle = \langle Z_j, \gamma_j\rangle_2 = \| {Z^T}_j\gamma_j \| \leq \lambda$; If $\hat{\theta}_j \neq 0$, ${-Z_j}^T(\gamma_j - Z_j\theta_j) + \lambda\frac{\theta_j}{\|\theta_j\|_2} = 0$, then $\hat{\theta}_j = ({Z^T}_j Z_j + \frac{\lambda}{\|\hat{\theta}_j\|})^{-1}{Z^T}_j\gamma_j$, a iteration process and we can use CCD algorithm.

  Under the orthonormality condition ($Z_j^T Z_j = 1$), $\arg\min_\theta f(\theta_j) \Rightarrow \arg\min_\theta \frac{1}{2}\| \gamma_j - Z_j\theta_{-j} \|^2 + \frac{1}{2}\| \theta_{-j} - \theta_j \|^2 + \lambda S_j \Rightarrow \arg\min_\theta \frac{1}{2}\| \theta_{-j} - \theta_j \|^2 + P_\lambda(\theta_j)$, where $\theta_{-j} = {Z^T}_j\gamma_j$. For LASSO($L_1$ penalty), the sol. of $\min_{\theta_j}\{\frac{1}{2}\| \theta_{-j} - \theta_j \|^2 + P_\lambda(\theta_j)\}$ satisfies: $\hat{\theta}_j(\theta_{-j}) = (1 - \frac{\lambda}{\hat{\theta}_{-j}})_+\hat{\theta}_{-j}$, i.e., $\hat{\theta}_j = (1 - \frac{\lambda}{\|{Z^T}_j\gamma_j\|_2})_+{Z^T}_j\gamma_j$. □

7. Elastic-net sol.

For the penalized least squares, the Elastic Net estimator is defined as

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\},$$

where $p_{\lambda_1,\lambda_2}(t) = \lambda_1|t| + \lambda_2 t^2$ is called the Elastic Net penalty. Another form of the Elastic Net penalty is

$$p_{\lambda,\alpha}(t) = \lambda J(t) = \lambda \left[(1-\alpha)t^2 + \alpha|t| \right],$$

with $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$. The Elastic Net is a pure ridge regression when $\alpha = 0$ and a pure Lasso when $\alpha = 1$. For the sol. of Elastic-net, we can differentiate the function $f(\beta) = \frac{1}{2}\| Y - X_j\beta_j \|^2_2 + \lambda[\frac{1}{2}(1-\alpha)\| \beta \|^2_2 + \alpha\| \beta \|_1]$ w.r.t. β_j and set it to zero to find out the sol.

$$\begin{aligned} -X_j(Y - X\beta) + \lambda(1-\alpha) + \lambda\alpha S(\beta_j) = -X_j(\gamma_j - X\beta_j) + \lambda(1-\alpha)\beta_j + \lambda\alpha S(\beta_j) = 0 \\ \Rightarrow \lambda(1-\alpha)\beta_j = X_j\gamma_j - \lambda\alpha S(\beta_j) \end{aligned}$$

The subgradient equation assure that

$$\hat{\beta}_j = \frac{S_{\lambda\alpha}(X_j\gamma_j)}{X_j\gamma_j + \lambda(1-\alpha)} \quad \square$$

8. Fused-lasso signal approximator

- (a.)

$$-2\sum_{i=1}^{N}(Y_i - \theta_0 - \theta_i) = 0 \Rightarrow \hat{\theta}_0 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\theta}_i)$$

- (b.)

Sol. of SCAD under Orthonormal design

We have notations: LSE of parameter, $z = \beta_{ols} = \frac{1}{n}\langle X, Y\rangle$(orthogonal case); Penalty function, $P_\lambda(\beta)$. Firstly, note that the LASSO sol. $\hat{\beta}(z) = sgn(z)(|z| - \lambda)_+$, which can be inducted by subgradient and is a $L_q(q = 1)$ penalty.

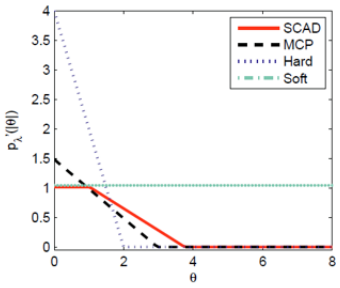
$$Slove: LASSO \ \arg\min_{\beta} \frac{1}{2n} \| Y - X\beta \|_2{}^2 + \| \beta \|_1 \Rightarrow z = sgn(\beta)(|\beta| + \lambda) \Rightarrow \hat{\beta}(z) = ? \quad (5)$$

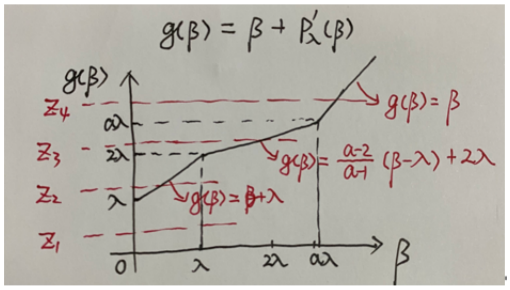
- if $\beta > 0$, $\beta + \lambda - z = 0$, we have $\hat{\beta} = z - \lambda$ under $z - \lambda > 0$, i.e., when $z > \lambda$, $\hat{\beta} = z - \lambda$.
- if $\beta = 0$, $sgn(\lambda) = z$, we have $|z| < \lambda$, i.e., when $|z| < \lambda$, $\hat{\beta} = 0$.
- if $\beta < 0$, $\beta - \lambda - z = 0$, we have $\hat{\beta} = z + \lambda$ under $z + \lambda < 0$, i.e., when $z < -\lambda$, $\hat{\beta} = z + \lambda$.

Secondly, for the SCAD sol., we can do some inductions under orthonormal design.

$$Slove: SCAD \ \arg\min_{\beta} \frac{1}{2} \| z - \beta \|_2{}^2 + \| P_\lambda(|\beta|) \|_1 \Rightarrow \arg\min_{\beta} \frac{1}{2}(z - \beta)^2 + P_\lambda(|\beta|) \Rightarrow \hat{\beta}(z) = ? \quad (6)$$

where we can assure that $\hat{\beta}(z) = -\hat{\beta}(-z)$ and $\hat{\beta}(z) \leq 0$ when $z \leq 0$ holds. So we only need to calculate $\hat{\beta}(z)$ when $z \leq 0$, i.e., $f_z(\beta) = \beta + P^{'}_{\lambda}(\beta) - z = 0, z \leq 0$.





- if $z \in [0, \lambda)$, $g(\beta) \geq \lambda > z$, i.e., $f_z(\beta) \geq 0$, $\hat{\beta(z)} = 0$.
- if $z \in [\lambda, 2\lambda)$,$g(\beta) = \beta + \lambda$, $\hat{\beta(z)} = z - \lambda$.
- if $z \in [2\lambda, a\lambda)$, $g(\beta) = \frac{a-2}{a-1}(\beta - \lambda) + 2\lambda$, $\hat{\beta(z)} = \frac{1}{a-2}[(a-1)z - a\lambda]$.
- if $z \in [a\lambda, +\infty)$, $g(\beta) = \beta$, $\hat{\beta(z)} = z$.