

401-3627-00: High-Dimensional Statistics

Conceptual Summary

Swiss Federal Institute of Technology Zurich (ETH Zürich)

Last Edit Date: 09/09/2025

Disclaimer and Term of Use:

1. We do not guarantee the accuracy and completeness of the summary content. Some of the course material may not be included, and some of the content in the summary may not be correct. You should use this file properly and legally. We are not responsible for any results from using this file.
2. This personal note is adapted (derived) from the slides and codes from *Professor Peter Bühlmann*. Please [contact us](#) to delete this file if you think your rights have been violated.
3. This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

1. Linear Models

$Y_i = \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i$

$i=1, \dots, n$ Matrix form

ϵ_i is i.i.d.

$E[\epsilon_i] = 0$

Unusual aspect: $p \gg n$

2. The Lasso Estimator

① Drawbacks of OLS

When $p \gg n$, OLS result is not unique, and will overfit the data, residual $\rightarrow 0$

③ Ridge - l_2 : $\hat{\beta}(\lambda) = \arg\min (\|Y - X\beta\|_2^2/n + \lambda \|\beta\|_2^2)$

3. Orthogonal design

If $p \leq n$, then design matrix $X^T X/n = I_{p \times p}$

① Lasso solution $\hat{\beta}_L(\lambda) = \text{sign}(Z_j)(|Z_j| - \lambda)_+$ has less bias

where $Z_j = (X^T Y)_j/n = \beta_{OLS}$

Soft $\beta_{OLS} = X^T Y/n$

Hard-threshold $\hat{\beta}_H(\lambda) = \text{sign}(Z_j)(|Z_j| - \lambda)_+$

② Lasso bias $E[\hat{\beta}_L(\lambda)] \neq \beta$ hard to compute!

④ Hard-threshold $\rightarrow$ $\| \cdot \|_0$ penalization

Simply $\hat{\beta}(\lambda) = \arg\min (\|Y - X\beta\|_2^2/n + \lambda \|\beta\|_0)$

Equals to β_{OLS} with JR , e.g. AIC, BIC, ~~non-convex~~

⑤ Implications and Asymptotic Theorem

$\|X(\hat{\beta} - \beta^*)\|_2^2/n = O_p(\lambda \|\beta^*\|_1)$ ~~logarithmic~~

Even for very sparse case with $\|\beta^*\|_1 = O(1)$, slow convergence rate of order $O_p(\log(p)/n)$.

No assumption on X for slow rate.

⑥ Lasso Properties and Sufficient Conditions to Address them

property	design condition	size of non-zero coeff.
slow prediction conv. rate	no requirement	no requirement
fast prediction conv. rate	compatibility	no requirement
estimation error bound $\ \hat{\beta} - \beta^* \ _1$	compatibility	no requirement
variable screening	or restricted eigenvalue	weaker beta-min cond.
variable selection	neighborhood stability	beta-min condition
	$\Leftrightarrow$ irreparable cond.	

Lasso identifies β^* under weaker conditions - ~~same condition~~

4. Consistency, conditions, convergence of Lasso

① Asymptotic Results

- Fixed design if $\|\beta^*\|_1 = O(\sqrt{n \log(p)})$, then $\|X(\hat{\beta} - \beta^*)\|_2^2/n = O_p(1)$, slow rate, consistency

- Fixed design with a "comparing condition" i.e. restricted eigenvalue with $\delta_0 > 0$, then $\|X(\hat{\beta} - \beta^*)\|_2^2/n = O_p(\frac{\log(p)}{n})$, $S_0 = \{j: \beta_j^* \neq 0\}$

$\|\hat{\beta} - \beta^*\|_1 = O_p(\sqrt{\frac{\log(p)}{n}})$

② Basic Inequality Lemma

$\|X(\hat{\beta} - \beta^*)\|_2^2/n + \lambda \|\hat{\beta} - \beta^*\|_1 \leq \epsilon^T X(\hat{\beta} - \beta^*)/n + \lambda \|\hat{\beta} - \beta^*\|_1$

1. If $S_0 = \{j: \beta_j^* \neq 0\}$, slow convergence rate $\|X(\hat{\beta} - \beta^*)\|_2^2/n = O_p(\frac{\log(p)}{n}) \rightarrow 0$ ($p, n \rightarrow \infty$)

2. If $S_0 = \{j: \beta_j^* \neq 0\}$, fast convergence rate $\|X(\hat{\beta} - \beta^*)\|_2^2/n = O_p(\sqrt{\frac{\log(p)}{n}})$ ($p, n \rightarrow \infty$)

③ Probabilistic Lemma

Assume $\sigma^2 = 1$ for all j . Then we have for all $t \geq 0$ and for $\lambda_0 = 2\sigma \sqrt{\frac{\log(p)}{n}}$

$P(|T| \geq t) \leq 2 \exp(-t^2/2)$

Assumes we know $\sigma^2 = \text{Var}(\epsilon_i)$ slow rate

④ Consistency of the Lasso

Assume $\hat{\sigma}^2 = 1$ for all j . For some $t > 0$, let the regularization parameter be $\lambda = 4\sigma \sqrt{\frac{\log(p)}{n}}$, where $\hat{\sigma}$ is some estimator of σ . With probability at least $1 - \alpha$, where $\alpha = 2 \exp(-t^2/2) + P(|\hat{\sigma} - \sigma| \geq t)$, we have $2\|X(\hat{\beta} - \beta^*)\|_2^2/n \leq 3\lambda \|\hat{\beta} - \beta^*\|_1$

⑤ Oracle Inequality

Assume compatibility condition & constant $\phi_0^2 \geq \lambda_0$ hold. On T and for $\lambda \geq 2\lambda_0$:

$\|X(\hat{\beta} - \beta^*)\|_2^2/n + \lambda \|\hat{\beta} - \beta^*\|_1 \leq 4\epsilon^T \beta^*/n$

When does this hold?

X is iid from sub-Gaussian distribution with $\lambda_0 = 0$. Z_i . Assume $\liminf(\Sigma) > 0$, $S_0 = \{j: \beta_j^* \neq 0\}$. Then, for $C > 0$, $\phi_0^2 \geq C \lambda_{\min}^2(\Sigma)$ with high probability.

- Estimation error for $\hat{\beta}^*$ in terms of l_1 norm.

5. Adaptive Lasso

Two stages, runs Lasso twice.

Good way to address bias of Lasso for orthogonal design.

Step 1: Initial estimator $\hat{\beta}_{init}$, e.g. the Lasso.

Step 2: Reweighed l_1 -penalty (each variable has its penalty)

$\hat{\beta}_{adapt}(\lambda) = \arg\min (\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_{init,j}|})$

Works well in practice. More sparse than Lasso. Better variable screening & selection if truth is assumed sparse.

Alternatives: thresholding Lasso, relaxed Lasso.

Computationally more complicated. Differences in large scale data is not obvious.

6. Computation Algorithm for Lasso (KKT)

As $\|\beta\|_1$ is not differentiable we use Karush-Kuhn-Tucker (KKT) conditions by taking the sub-differential: $\frac{\partial}{\partial \beta} (\|Y - X\beta\|_2^2/n + \lambda \|\beta\|_1) = G_j(\beta) + \lambda e_j$

$G_j(\beta) = -2X^T(Y - X\beta)/n$, $e_j = \text{sign}(\beta_j)$ if $\beta_j \neq 0$, $e_j \in [-1, 1]$ if $\beta_j = 0$.

KKT condition: $G_j(\beta) = -\text{sign}(\beta_j)\lambda$, if $\beta_j \neq 0$; $|G_j(\beta)| \leq \lambda$, if $\beta_j = 0$. Algorithm (simplified):

- Let $\beta^{(0)} \in \mathbb{R}^p$ be an initial parameter vector. For $m = 1, 2, \dots$
- repeat
- Proceed componentwise $j = 1, 2, \dots, p$, $1, 2, \dots, p, 1, 2, \dots$
- update: if $|G_j(\beta^{(m-1)})| \leq \lambda$: set $\beta_j^{(m)} = 0$, prev. parameter with j th comp $= 0$ otherwise: $\beta_j^{(m)}$ is the minimizer of the objective function with respect to the j th component but keeping all others fixed
- until numerical convergence

7. Lasso Regularization Path

If $\lambda_{max} = 2X^T Y/n$ then $\hat{\beta}(\lambda_{max}) = 0$. Not monotone in the non-zeros. $\hat{\beta}_j(\lambda) \neq 0 \Rightarrow \hat{\beta}_j(\lambda) = 0, \lambda' < \lambda$

8. Generalized Linear Models

Univariate Y , Covariate $X \in \mathbb{R}^p$

$Y_1, \dots, Y_n$ independent

$g(E[Y_i|X_i = x]) = \mu + \sum_{j=1}^p \beta_j x_j$

conditional mean = fix = $f_{\beta}(x)$

- $g(\cdot)$ real-valued link function

unknown link function. Lasso - l_1 -norm

- μ an intercept term. Negative log-likelihood

9. Group Lasso

$Y = X\beta + \epsilon$, β (high-dim) param vector

$\beta = (\beta_{g_1}, \dots, \beta_{g_q})$, $\beta_{g_j} = \{\beta_r: r \in g_j\}$, $\bigcup_{j=1}^q g_j = \{1, \dots, p\}$

$g_j \cap g_k = \emptyset$ ($j \neq k$)

① Example $Y \in \mathbb{R}$, $X = (X^{(1)}, \dots, X^{(p)})$, $Y_i, X_i^{(k)} \in \{0, 1\}$

$p=2$, then $Y_i = \mu + \sum_{k=1}^2 \beta_k I(X_i^{(k)} = 1) + \sum_{k=1}^2 \beta_k I(X_i^{(k)} = 0)$ main effects

interaction effects $+ \sum_{k=1}^2 \beta_k I(X_i^{(1)} = 1, X_i^{(2)} = 1) + \epsilon_i$

Constraints: $(1+1+4+3)=9$ Params: $(1+4+4+16)=9$ free params.

② Group Lasso Penalty want either $\beta_{g_j} = 0$ or $(\beta_{g_j})_r \neq 0 \forall r \in g_j$

$\hat{\beta}(\lambda) = \arg\min (\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^q \|\beta_{g_j}\|_2)$, $m_j = \sqrt{|g_j|}$

$m_j \|\beta_{g_j}\|_2 \leq |g_j|$ if $(\beta_{g_j})_r \neq 0 \forall r$, $\sum_{r \in g_j} (\beta_{g_j})_r \leq |g_j|$ if $(\beta_{g_j})_r \neq 0 \forall r$

Sparse happens at points/regions of non-differentiability. Similar to ridge

$\|\beta_{g_j}\|_2 = \sqrt{\|\beta_{g_j}\|_1^2}$, $\frac{\partial}{\partial \beta_{g_j}} = \frac{1}{2} (\|\beta_{g_j}\|_1)^{-1/2} \cdot 2\beta_{g_j} = \beta_{g_j} / \|\beta_{g_j}\|_2$ ridge

If $\|\beta_{g_j}\|_2 = 0 \Leftrightarrow \beta_{g_j} = 0$ point of non-differentiability condition satisfied

10. Additive Models

$Y_i = \mu + \sum_{j=1}^p f_j(X_i^{(j)}) + \epsilon_i$, $i=1, \dots, n$

$E[\epsilon_i] = 0$, ϵ_i independent of X_i , X_i random

Denote by $(H_j)_{j=1, \dots, p} = (h_{j,k}(X_i^{(j)}))_{k=1, \dots, p}$, $\beta_j = (\beta_{j,1}, \dots, \beta_{j,p})^T$

$f_j(\cdot): \mathbb{R} \rightarrow \mathbb{R}$, smooth

$p \times 1 \Rightarrow \beta = (\beta_1, \beta_2, \dots, \beta_p)^T \Rightarrow (\sum_{j=1}^p f_j(X_i^{(j)}))_{i=1}^n = \sum_{j=1}^p H_j \beta_j$

$\Rightarrow \hat{\beta}_j = 0$ or $\hat{\beta}_j \neq 0$ with $\hat{\beta}_{j,k} \neq 0 \forall k$

① Group Lasso Type Problem $Y_i = \mu + \sum_{j=1}^p f_j(X_i^{(j)}) + \epsilon_i$

$f_j(\cdot) = \sum_{k=1}^p \beta_{j,k} h_{j,k}(\cdot) \rightarrow$ Basis func.: splines, Fourier waves

② Sparse Additive Models (SpAM)

$\hat{\beta} = \arg\min (\|Y - \sum_{j=1}^p H_j \beta_j\|_2^2/n + \text{pen}_\lambda(\beta))$ Groupwise $\rightarrow$ prediction

$\text{pen}_\lambda(\beta) = \lambda \sum_{j=1}^p \|H_j \beta_j\|_2 / \sqrt{n} = \lambda \sum_{j=1}^p \|f_j\|_2$ penalty

$\Rightarrow \lambda \sum_{j=1}^p \sqrt{\beta_j^T H_j^T H_j \beta_j} / \sqrt{n}$ K is small, H_j positive definite

SpAM does not explicitly involve smoothness of $f(\cdot)$. It is computationally impossible to choose different K_j for every $f_j(\cdot)$. Splines take smoothness into account.

③ Natural cubic splines & Sobolev spaces

$\hat{f}_1, \dots, \hat{f}_p = \arg\min_{f_1, \dots, f_p} (\|Y - \sum_{j=1}^p f_j\|_2^2/n + \lambda \sum_{j=1}^p \|f_j\|_2 + \lambda \sum_{j=1}^p \int |f_j''(x)|^2 dx)$

$f_j: [a, b] \rightarrow \mathbb{R}$; twice continuously differentiable and $\int_a^b |f_j''(x)|^2 dx < \infty$ $\rightarrow$ $I^2(f)$ smoothness index measure of roughness.

