TOYOHASHI
UNIVERSITY OF TECHNOLOGY

ISIT JUNE 22-27 ANN ARBOR MICHIGAN

Empirical Bayes Estimation for Lasso-Type Regularizers

— Analysis of Automatic Relevance Determination —

Tsukasa Yoshida, Kazuho Watanabe

Profiles

Tsukasa Yoshida
PhD Student, Toyohashi University / NTT
Japan

Kazuho Watanabe
Professor, Toyohashi University
Japan

Summary

- Empirical Bayes (EB) estimation of hyperparameters in Ridge, Lasso, or Group Lasso **seems to** induce automatic relevance determination (ARD).

including a conjecture

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \lambda h(\boldsymbol{w})$$

hyperparameter

EB for λ

$$\hat{\lambda} = \begin{cases} \lambda^* & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 > \|\Phi\|_F\right) \\ \infty & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 \leq \|\Phi\|_F\right) \end{cases}$$

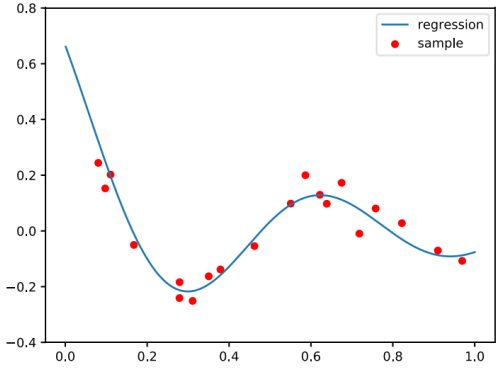
This means ARD occurs!

Background

Our Focus

◆ We focus on linear regression models including Ridge, Lasso, and Group Lasso.

$$\underbrace{L(\boldsymbol{w})}_{\text{Loss function}} = \underbrace{\frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2}_{\text{Error term}} + \underbrace{\lambda h(\boldsymbol{w})}_{\textbf{Regularizer}}$$



Examples:

- $h(w) = \frac{1}{2}||w||_2^2 \rightarrow$ Ridge: $L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$
- $h(w) = ||w||_1 \rightarrow$ Lasso: $L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$
- $h(w) = ||w||_2 \rightarrow$ Group Lasso: $L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \sum_{g\in\mathfrak{G}} \lambda_g \|\boldsymbol{w}_g\|_2$

Problem to Address

◆ We want to determine the hyperparameter automatically.

Before estimating the main parameter w based on the loss function, it is necessary to set the hyperparameters λ.

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \lambda h(\boldsymbol{w})$$

parameter　　hyperparameter

Estimation methods

◆ Mainly cross-validation and empirical Bayes

Cross-validation (CV):

Method for selecting hyperparameters by evaluating model performance across multiple train-validation splits of the data

Empirical Bayes (EB):

Method for determining hyperparameters by $\hat{\lambda} := \arg\max_{\lambda} Z(\lambda)$,

where $Z(\lambda) = \int p(y^n \mid \boldsymbol{x}^n, \boldsymbol{w}) p(\boldsymbol{w} \mid \lambda) d\boldsymbol{w}$

Estimation methods

◆ Mainly cross-validation and empirical Bayes

Cross-validation (CV):

Method for selecting hyperparameters by evaluating model performance across multiple train-validation splits of the data

Empirical Bayes (EB):

Method for determining hyperparameters by $\hat{\lambda} := \arg\max_{\lambda} Z(\lambda)$,

where $Z(\lambda) = \int p(y^n \mid \boldsymbol{x}^n, \boldsymbol{w}) p(\boldsymbol{w} \mid \lambda) d\boldsymbol{w}$

Our focus

Difficulty of EB

◆ Analysis of the marginal likelihood is challenging.

Problem to solve:

$$\hat{\lambda} := \arg\max_{\lambda} Z(\lambda)$$

Calculation of the marginal likelihood Z:

$$Z(\lambda) = \int p(y^n \mid \boldsymbol{x}^n, \boldsymbol{w}) p(\boldsymbol{w} \mid \lambda) d\boldsymbol{w} \propto \int e^{-L(\boldsymbol{w})} d\boldsymbol{w}$$

This integration is very difficult.

Property of EB

◆ In Ridge regression, it is known that EB induces ARD.

EB estimator for Ridge:

$$\hat{\lambda}_{\mathrm{ridge}} = \begin{cases} \dfrac{mn}{\|\tilde{\boldsymbol{y}}\|_2^2 - m} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \leq \sqrt{m}) \end{cases}$$

Estimation of parameter w when $\hat{\lambda} = \infty$:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\ &= \boldsymbol{0} \end{aligned}$$

← This can be used for sparse parameter selection.

Property of EB

- In Ridge regression, it is known that EB induces ARD.

EB estimator for Ridge:

$$\hat{\lambda}_{\mathrm{ridge}} = \begin{cases} \dfrac{mn}{\|\tilde{\boldsymbol{y}}\|_2^2 - m} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \leq \sqrt{m}) \end{cases}$$

For simplicity, we refer to the divergence of the estimator of λ as ARD.

Estimation of parameter w when $\hat{\lambda} = \infty$:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\ &= \boldsymbol{0} \end{aligned}$$

← This can be used for sparse parameter selection.

Recent Studies on EB

◆ ARD also occurs in Lasso, in addition to Ridge.

EB estimator for Ridge:

$$\hat{\lambda}_{\text{ridge}} = \begin{cases} \dfrac{mn}{\|\tilde{\boldsymbol{y}}\|_2^2 - m} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \leq \sqrt{m}) \end{cases}$$

Both can diverge to infinity!

EB estimator for Lasso:

$$\hat{\lambda}_{\text{lasso}} = \begin{cases} \lambda^*_{\text{lasso}} & (\text{if } |\tilde{y}| > 1) \\ \infty & (\text{if } |\tilde{y}| \leq 1) \end{cases}$$

Both derivations involve assumptions or restrictions on the parameters, such as w and the design matrix Φ. Therefore, in fact, these problems have not been completely solved.

(Supplement) Statstic Aspect of EB

◆ EB is important from a statistical perspective.

A major issue in statistics: How should we determine the prior distribution?

→ EB corresponds to a method for estimating the parameters of the prior distribution.

CV and EB have different statistical characteristics:

- CV corresponds to AIC,
- EB corresponds to BIC.

Our Results

Result 1

- We rigorously derived the EB estimator for Group Lasso.
  → ARD also occurs in Group Lasso.

$$\hat{\lambda}_{\mathrm{GL}} = \begin{cases} \lambda^*_{\mathrm{GL}} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \leq \sqrt{m}) \end{cases}$$

ARD!

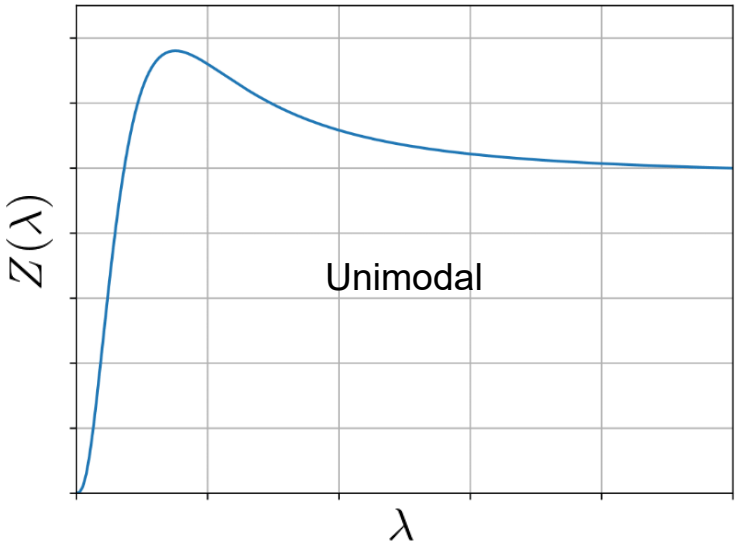
Note that this derivation also involves assumptions or restrictions on the parameters.

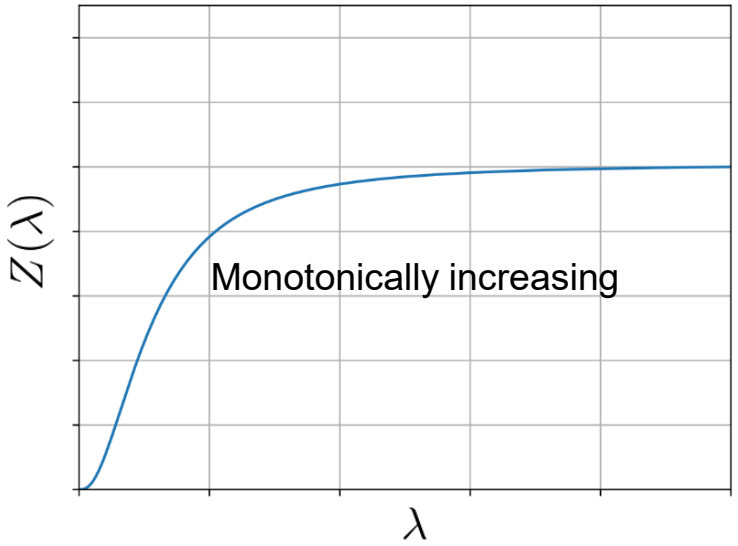
This is the third study in which the structure of ARD has been rigorously proven.

Key Part in the Proof of Result 1

- Quasiconcavity of the marginal likelihood Z

$$\hat{\lambda} := \arg\max_{\lambda} Z(\lambda) = \begin{cases} \lambda^*_{\mathrm{GL}} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \leq \sqrt{m}) \end{cases}$$





Question from Result 1 and Previous Studies

- EB induces ARD in Ridge, Lasso, and Group Lasso. How far can this be generalized?

$$\hat{\lambda}_{\mathrm{ridge}} = \begin{cases} \dfrac{mn}{\|\tilde{\boldsymbol{y}}\|_2^2 - m} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \le \sqrt{m}) \end{cases}$$

$$\hat{\lambda}_{\mathrm{lasso}} = \begin{cases} \lambda^*_{\mathrm{lasso}} & (\text{if } |\tilde{y}| > 1) \\ \infty & (\text{if } |\tilde{y}| \le 1) \end{cases}$$

$$\hat{\lambda}_{\mathrm{GL}} = \begin{cases} \lambda^*_{\mathrm{GL}} & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 > \sqrt{m}) \\ \infty & (\text{if } \|\tilde{\boldsymbol{y}}\|_2 \le \sqrt{m}) \end{cases}$$

Result 2

- We rigorously derived the EB estimator for linear regression with a homogeneous regularizer under the assumption that the Z is quasiconcave.

Loss function of linear regression with a homogeneous regularizer:

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \underbrace{\lambda h(\boldsymbol{w})}$$

This includes Ridge, Lasso, and Group Lasso.

EB estimator:

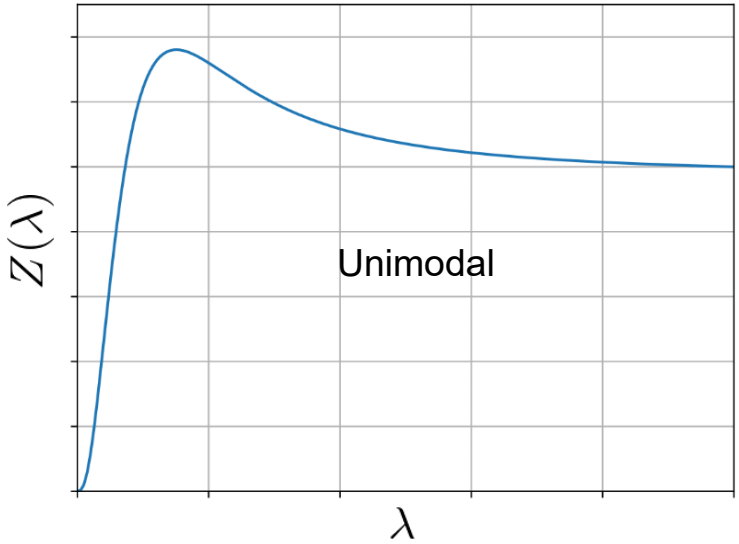
$$\hat{\lambda} = \begin{cases} \lambda^* & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 > \|\Phi\|_F\right) \\ \infty & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 \leq \|\Phi\|_F\right) \end{cases}$$

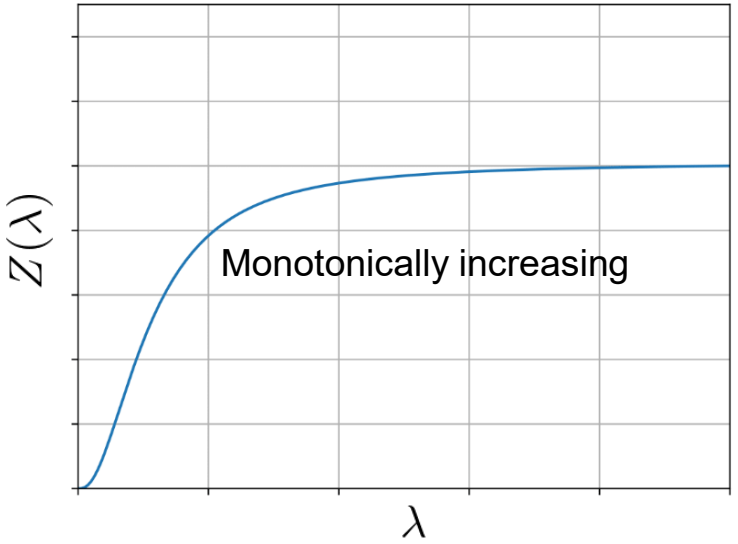
ARD!

Generalization of the conditions under which the estimator diverges

Question on the assumption in Result 2

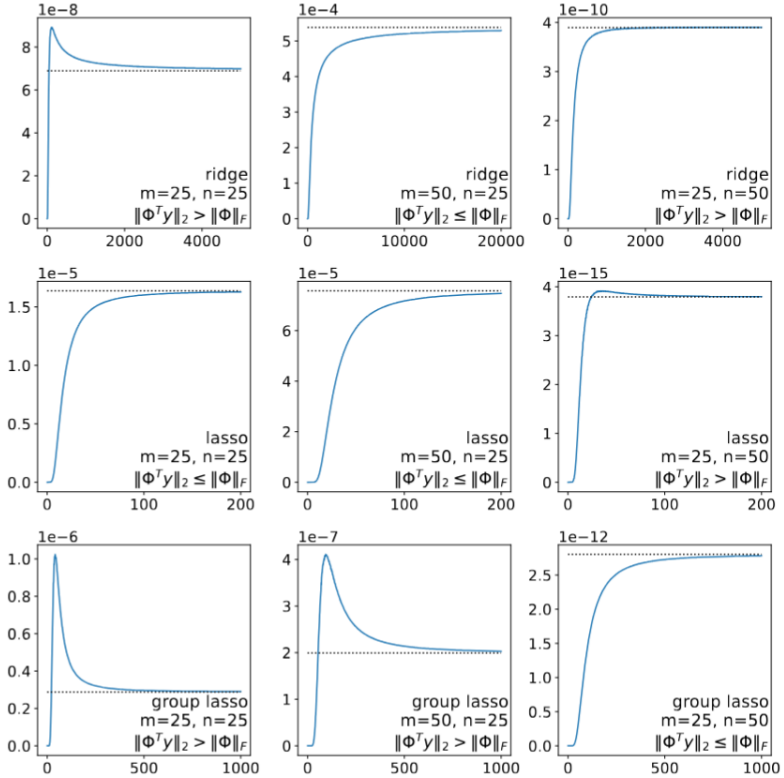
- Which models satisfy the quasiconcavity of Z?





Result 3

- We confirmed through numerical experiments that Z is quasiconcave in Ridge, Lasso, and Group Lasso.

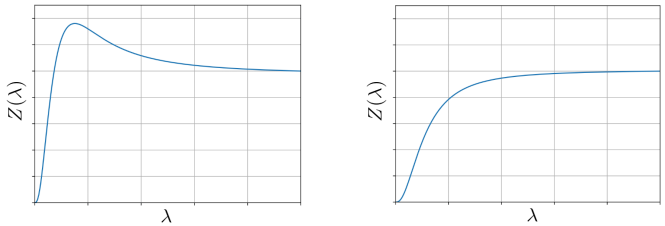


Conjecture from Result 3

- When using a homogeneous regularizer, such as in Ridge, Lasso, or Group Lasso, Z is likely to be quasiconcave.

Conjecture:

Under the loss function with a homogeneous regularizer, the marginal likelihood is quasiconcave.



If this conjecture is true, the EB estimator for λ is expressed as:

$$\hat{\lambda} = \begin{cases} \lambda^* & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 > \|\Phi\|_F\right) \\ \infty & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 \leq \|\Phi\|_F\right) \end{cases}$$

Conclusion

- Empirical Bayes (EB) estimation of hyperparameters in Ridge, Lasso, or Group Lasso **seems to** induce automatic relevance determination (ARD).

including a conjecture

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|_2^2 + \lambda h(\boldsymbol{w})$$

hyperparameter

EB for λ

$$\hat{\lambda} = \begin{cases} \lambda^* & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 > \|\Phi\|_F\right) \\ \infty & \left(\text{if } \|\Phi^{\mathrm{T}}\boldsymbol{y}\|_2 \leq \|\Phi\|_F\right) \end{cases}$$

This means ARD occurs!