Penalized Likelihood in Bioinformatics: Ridge, Lasso, EN, SCAD, MCP

Yu Ding

MD Anderson Cancer Center

April 1, 2025

Outline

1. Motivation
2. Five Methods: Ridge, Lasso, EN, SCAD, MCP
3. Group Selection
4. Practical Tips & Oracle Property
5. References and Q&A

High-Dimensional Bioinformatics

Key Challenges:

- $\mathbf{p} \gg \mathbf{n}$: Many features (genes, SNPs) but few samples.
- Classic MLE is prone to overfitting or even undefined (if $p > n$).
- We need methods that reduce variance, control complexity.

Examples:

- Microarray / RNA-Seq: tens of thousands of genes, 50–200 samples.
- GWAS: up to millions of SNPs, typically a few thousand samples.

General Penalized Likelihood Formulation

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}} \big\{ -\ell(\boldsymbol{\beta}) + \lambda\, \Omega(\boldsymbol{\beta}) \big\},$$

where

- $\ell(\boldsymbol{\beta})$ is the log-likelihood (linear, logistic, Cox, etc.),
- $\Omega(\boldsymbol{\beta})$ is a penalty function (e.g., $\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{\beta}\|_1$, SCAD),
- $\lambda \geq 0$ controls the trade-off between fit and penalty.

Common Goals:

- **Sparsity** (for variable selection).
- **Stability** (shrink correlated features).
- **Lower bias** on large signals (non-convex methods).

Common Penalties (Overview)

Forms of $\Omega(\theta)$:

- $\|\theta\|_2^2$ (Ridge)
- $\|\theta\|_1$ (Lasso)
- Combination: $\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2$ (Elastic Net)
- Non-convex: SCAD, MCP

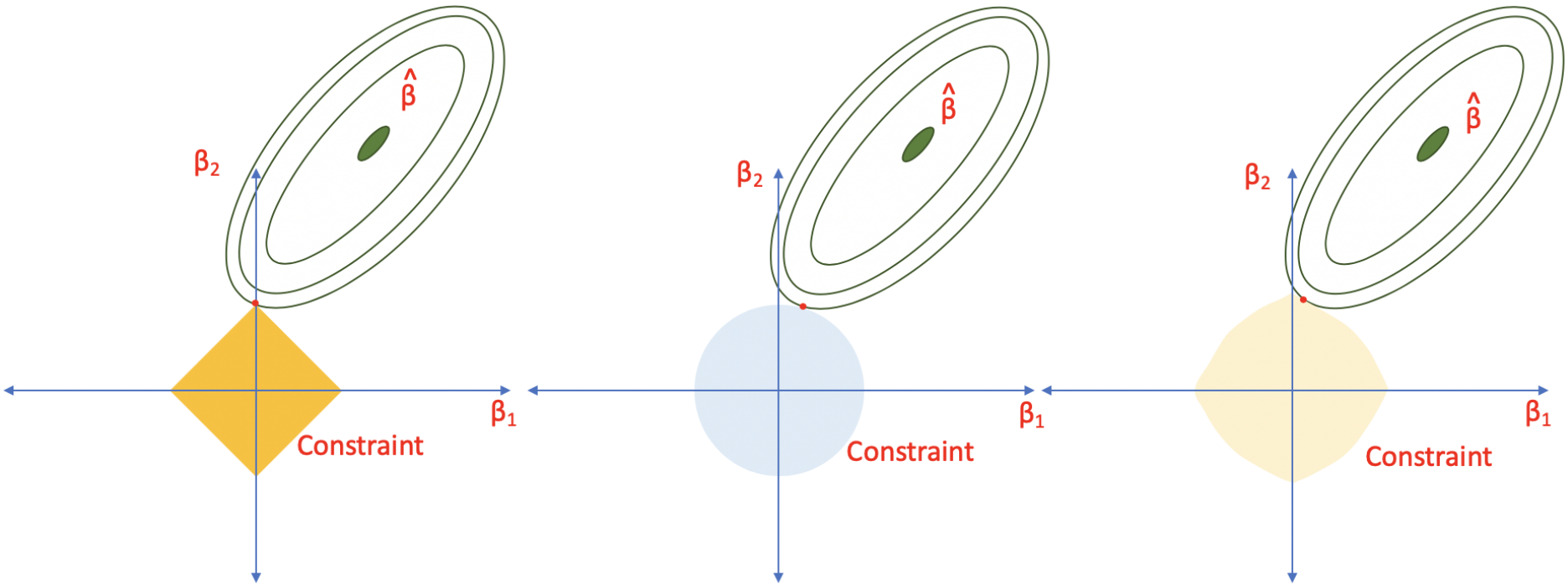
General Effects:

- L2 (Ridge): *continuous* shrinkage, no zero coefficients.
- L1 (Lasso): *sparsity*, some exact zeros.
- SCAD, MCP: *sparsity* + *less bias* on large coefficients.

Outline

1. Motivation
2. Five Methods: Ridge, Lasso, EN, SCAD, MCP
3. Group Selection
4. Practical Tips & Oracle Property
5. References and Q&A

Standard Penalties Overview



https://www.datasklr.com/extensions-of-ols-regression/
regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression

1. Ridge Regression

Reference (Method): Hoerl & Kennard (1970), *Technometrics*.
Penalty:

$$\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2.$$

Linear Version

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right].$$

Key Points:

- Coefficients shrink but seldom go to zero.
- Handles correlated features well (distributes weights).
- No direct feature selection (all remain in the model).

Ridge Example: Cell-Type–Specific DE

Reference (Example):

- Takeuchi & Kato (2021), *BMC Bioinformatics*.

Scenario:

- Bulk tissue expression with multiple cell types.
- Want to detect cell-type–trait interactions.
- High correlation in cell-type proportions (they sum to 1).

Nonlinear Ridge Approach: The method builds a nonlinear regression model that simultaneously analyzes two scales. In simplified form, if we denote for sample i:

- Y_i as the bulk omics measurement,
- W_{ih} as the proportion of cell type h,
- X_i as the trait (e.g., disease status or age),

Collinearity among interaction terms

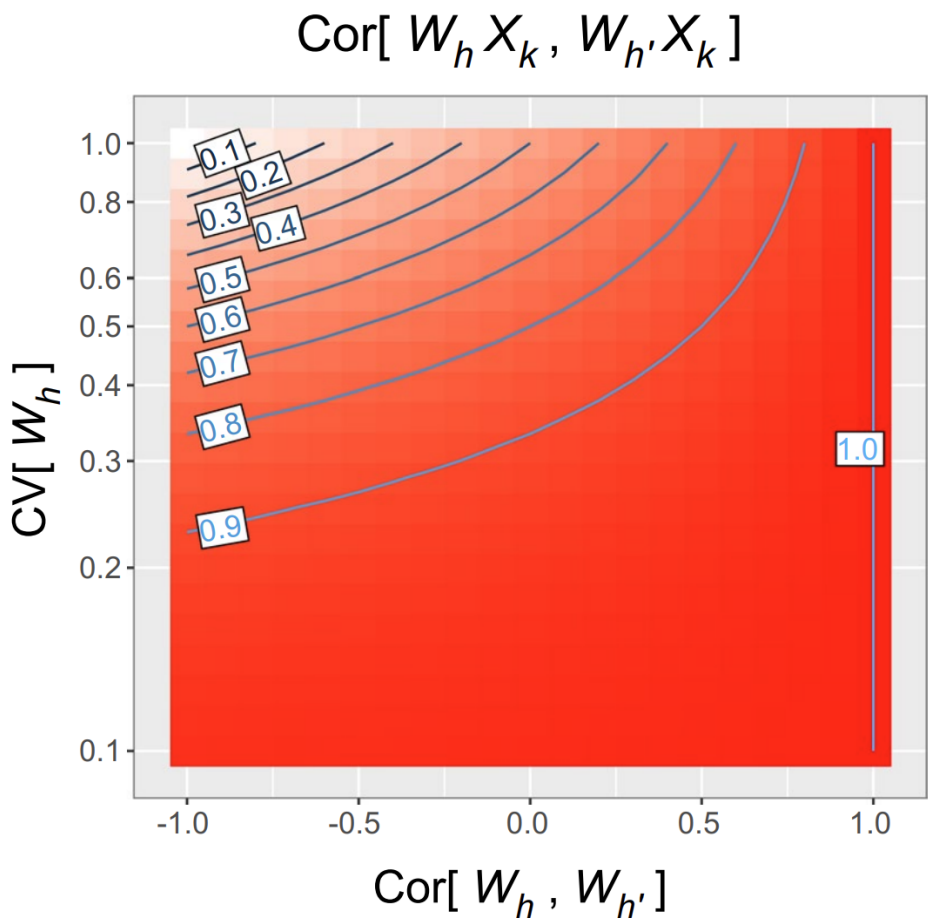


Fig. 1 Contour plot of the correlation coefficient between interaction terms $W_h X_k$ and $W_{h'} X_k$. W_h and $W_{h'}$ represent proportions of cell types h and h', and X_k represents the value of trait k. For this plot, we assume the coefficient of variation CV[W_h] and CV[$W_{h'}$] to be equal. As the CV decreases 0.6, 0.4 to 0.2, the correlation coefficient raises > 0.5, > 0.7 to > 0.9, over most range of Cor[W_h, $W_{h'}$]

Takeuchi & Kato (2021), *BMC Bioinformatics*

Nonlinear Ridge Regression for Cell-Type-Specific Analysis

$$Y_i = \sum_i f(W_{ih}, X_i) + \epsilon_i = \sum_h \beta_h (W_{ih} \times X_i) + \epsilon_i$$

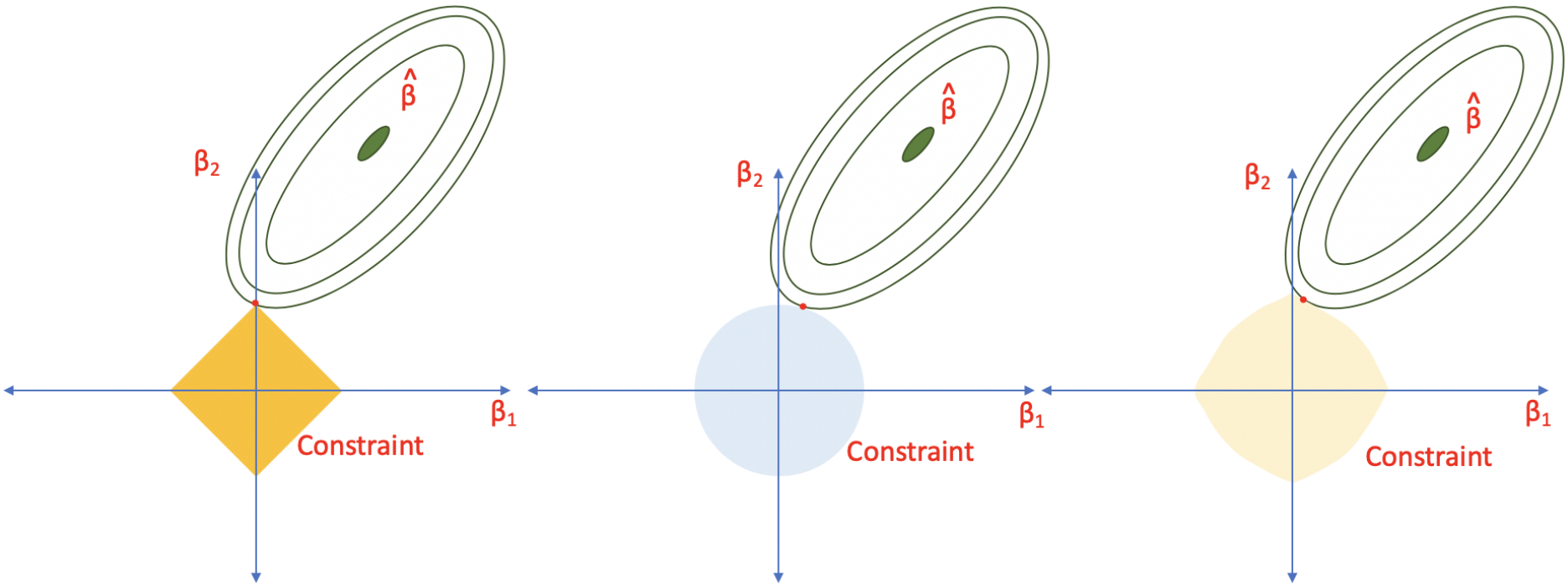
$$\ell = \sum_i (\hat{Y}_i - Y_i)^2 + \lambda \sum_h \beta_h^2$$

- Let $f(W_{ih}, X_i)$ be a function of cell proportion W_{ih} and trait X_i.
- Penalize with $\|\beta\|_2^2$ to avoid huge variance in correlated terms.

Outcome:

- Stable estimation of each cell type's effect.
- All cell types remain in the model (no zeros).

Standard Penalties Overview (Reminder)



https://www.datasklr.com/extensions-of-ols-regression/regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression

2. Lasso Regression

Reference (Method): Tibshirani (1996), *J. R. Stat. Soc. B.*
Penalty:

$$\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|.$$

Linear Version

$$\min_{\boldsymbol{\beta}} \Big[\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big].$$

Key Points:

- Creates **exact zeros** (feature selection).
- Potential over-shrinkage of large signals.
- Sensitive to correlated features (may pick one, drop others).

Lasso Example: Genome-wide Association Studies

Reference (Example):

- Wu et al. (2009), *Bioinformatics*.

Problem:

- Modern GWAS involve hundreds of thousands of SNPs ($p \gg n$).
- Univariate SNP-by-SNP tests ignore interactions and multi-collinearity.

Proposed Approach:

- **Lasso-penalized logistic regression** for joint variable selection.
- Zeroes out most SNP coefficients, retaining a small subset of putative causal variants.

GWAS with Lasso

Logistic Regression Model:

$$p_i = \frac{\exp(\mu + x_i^\top \beta)}{1 + \exp(\mu + x_i^\top \beta)}, \quad L(\theta) = \sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big],$$

$$g(\theta) = L(\theta) - \lambda \sum_{j=1}^{p} |\beta_j|, \quad \theta = (\mu, \beta_1, \ldots, \beta_p).$$

- Quickly drives most coefficients to zero.
- Simplifies interpretation in large-p settings.

An Example GWAS Plot

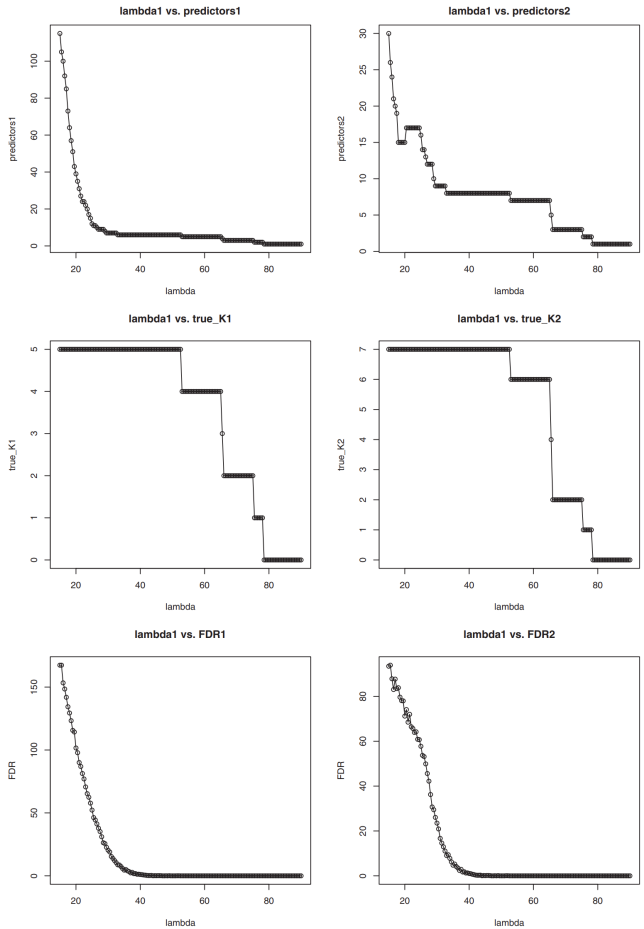
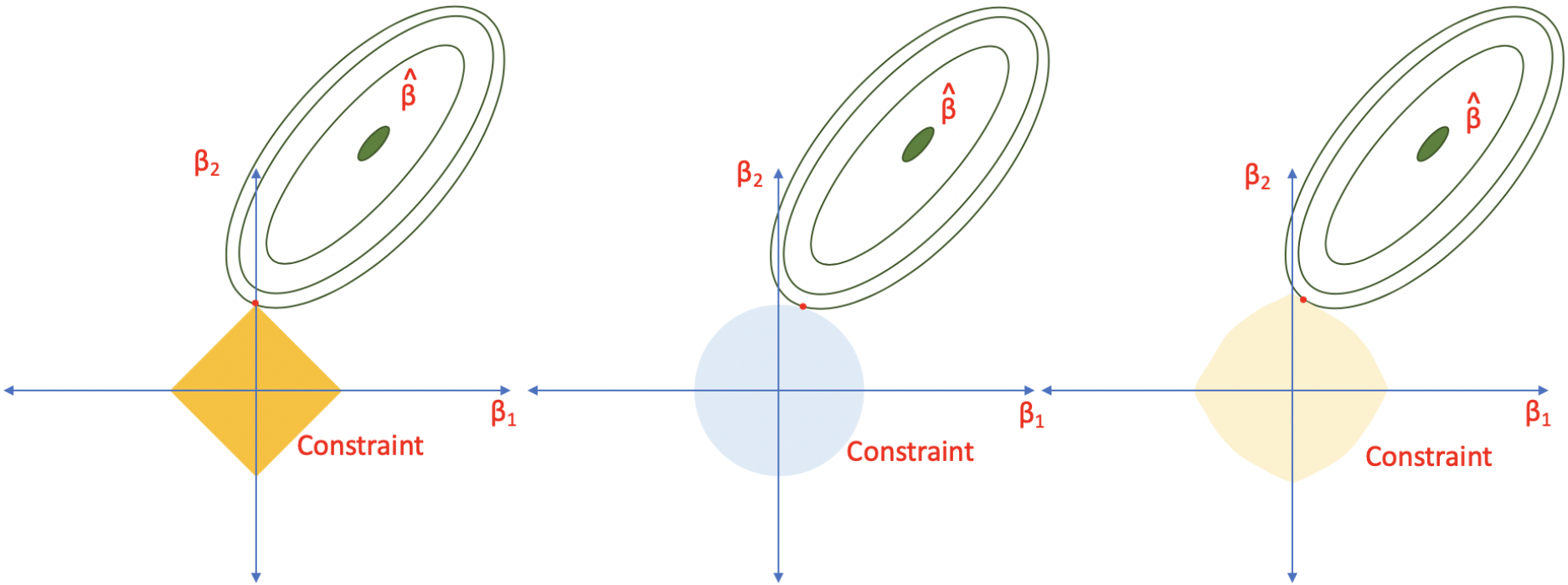


Fig. 1. Plots of the stage one penalty constant λ_1 versus the number of selected predictors, the number of true predictors and FDR. The stage two penalty constant $\lambda_2 = 25$.

Wu et al. (2009), *Bioinformatics*

Standard Penalties Overview (Reminder)



https://www.datasklr.com/extensions-of-ols-regression/
regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression

3. Elastic Net (EN)

Reference (Method): Zou & Hastie (2005), *J. R. Stat. Soc. B.*
Penalty (Blend of L1 & L2):

$$\Omega(\boldsymbol{\beta}) = \alpha \sum_j |\beta_j| + (1-\alpha) \sum_j \beta_j^2, \quad (0 \le \alpha \le 1).$$

Why use EN?

- More stable than Lasso under strong feature correlations.
- Still yields some zeros for variable selection.

An Elastic-Net–Based Regression Model (1/2)

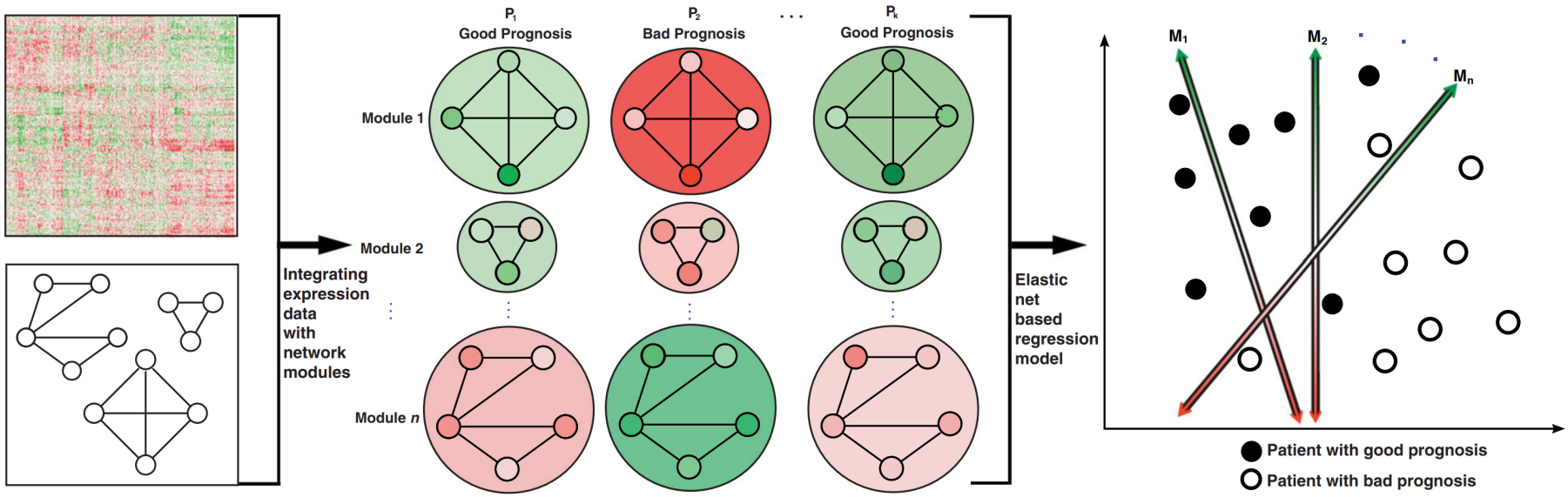
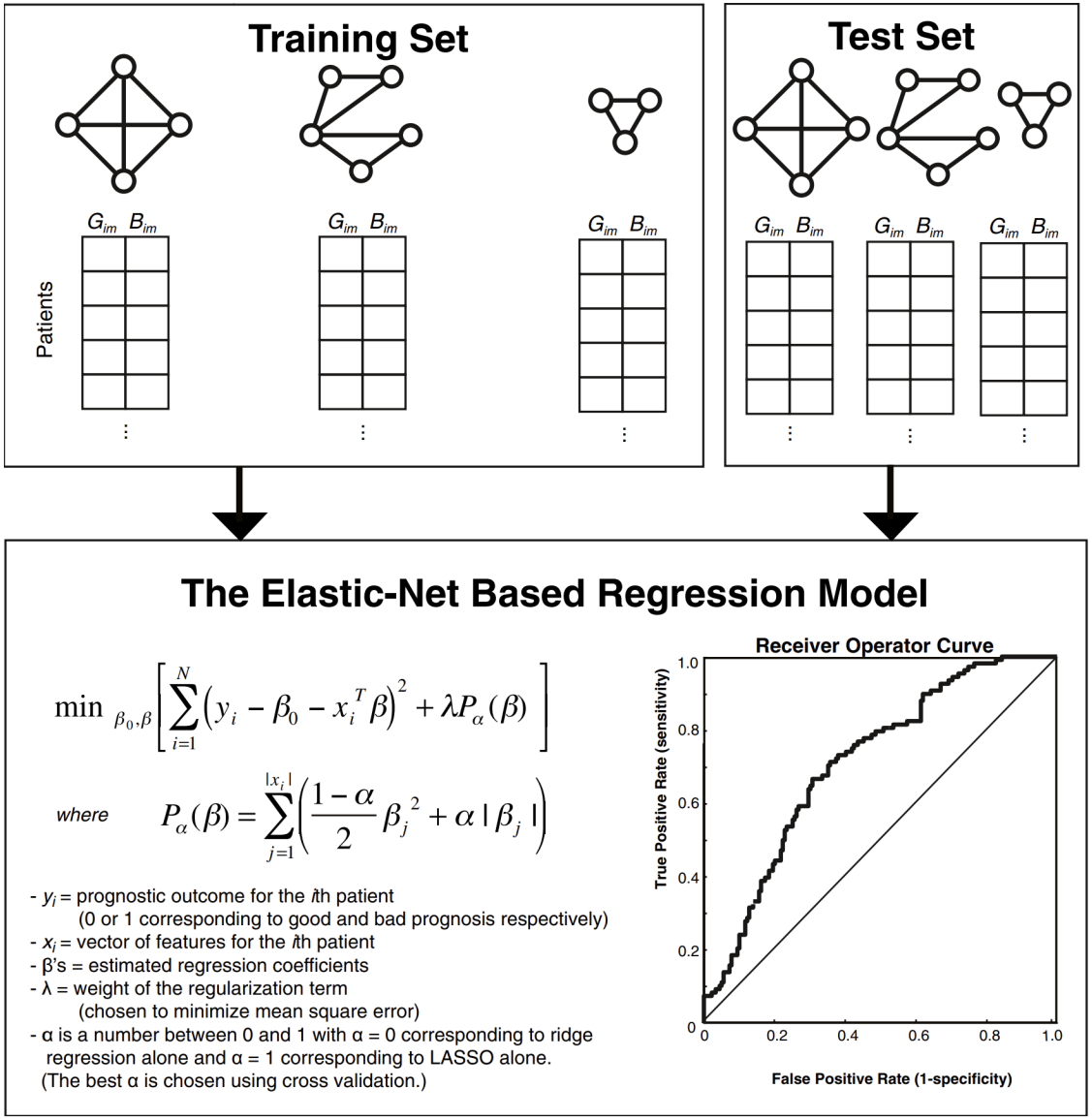


Figure 1 Schematic of ENCAPP. ENCAPP begins by overlaying tissue-specific gene expression data with the reference interactome network. Modules that have significant differential co-expression between patients with good and bad prognosis are used to build a regression model that can predict prognostic outcome.

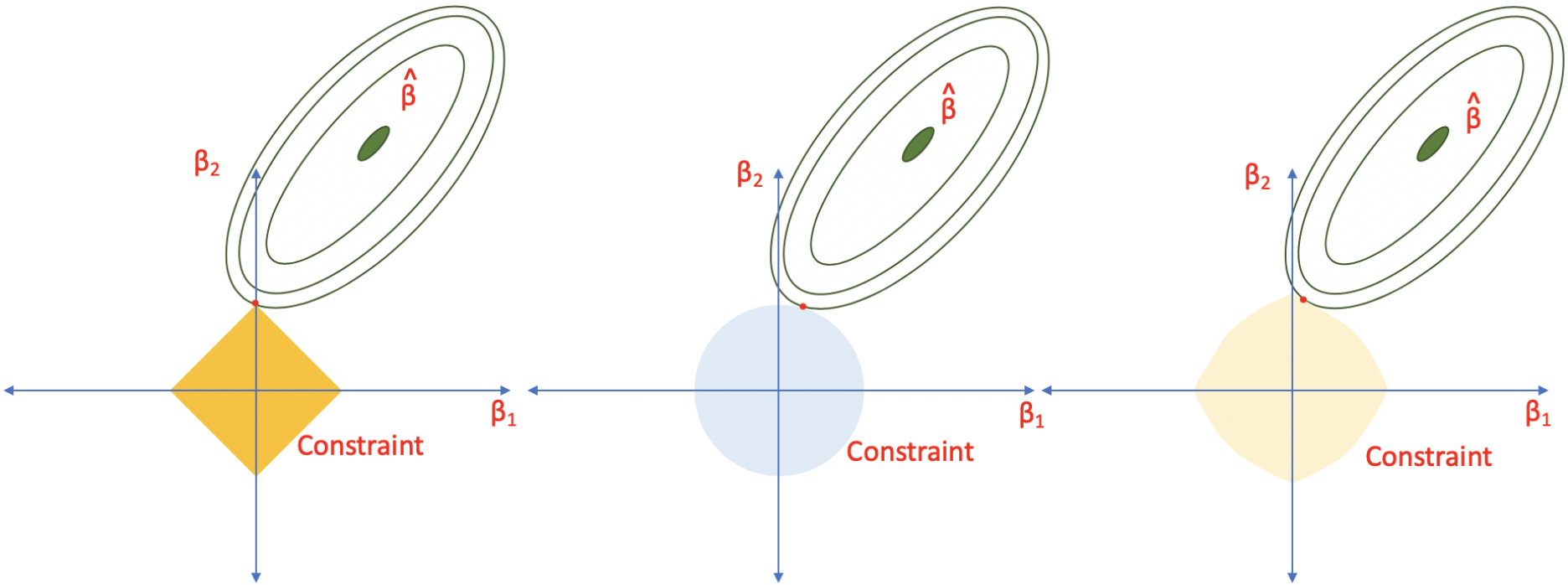
Das et al. (2015), *BMC Genomics*

An Elastic-Net–Based Regression Model (2/2)



Das et al. (2015), *BMC Genomics*

Standard Penalties Overview (Reminder)



https://www.datasklr.com/extensions-of-ols-regression/
regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression

4. SCAD (Smoothly Clipped Absolute Deviation)

Definition (Fan & Li, 2001)

$$P_\lambda^{\text{SCAD}}(\theta) = \begin{cases} \lambda|\theta|, & |\theta| \le \lambda, \\ -\frac{\theta^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}, & \lambda < |\theta| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |\theta| > a\lambda, \end{cases}$$

for $a > 2$ (often $a = 3.7$).

Pieces:

- L1-like near zero: ensures sparsity.
- Flatter penalty for large $|\theta|$: reduces bias.

Clonal structure identification through penalizing pairwise differences

A unsupervised learning for homogeneity pursuit

- mutation i, variant reads $r_i \sim \text{Binomial}(n_i, \theta_i)$
- total reads $n_i \sim \text{Poisson}(D)$
- variant allele fraction $\theta_i = \frac{\phi_i b_i}{(1-\rho)c_i^N + \rho c_i^T}$, where ϕ_i is cellular prevalence for mutation i

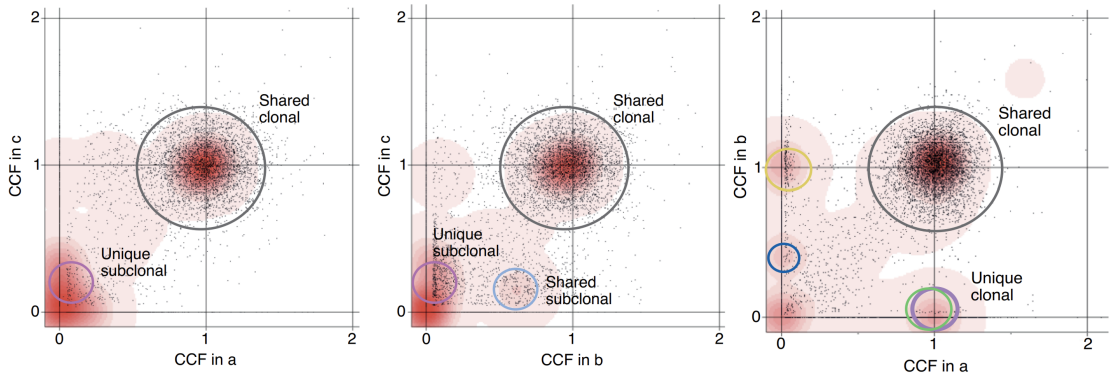
The log-likelihood for cellular prevalence

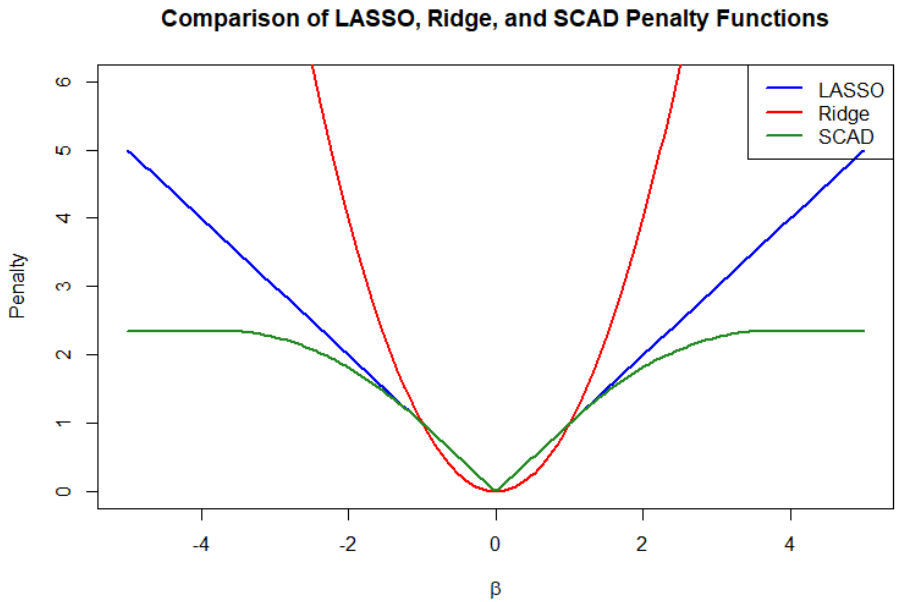
- $\ell(\phi) = \sum_i (r_i \log(\theta_i(\phi_i)) + (n_i - r_i) \log(1 - \theta_i(\phi_i)))$

Why use SCAD?

- Penalizing pairwise differences in ϕ_i to cluster similar mutations.
- SCAD's less aggressive shrinkage on large signals helps identify truly distinct clonal groups.

Clonal structure identification through penalizing pairwise differences





5. MCP (Minimax Concave Penalty)

Reference (Method): Zhang (2010), *Annals of Statistics*.
Key Idea:

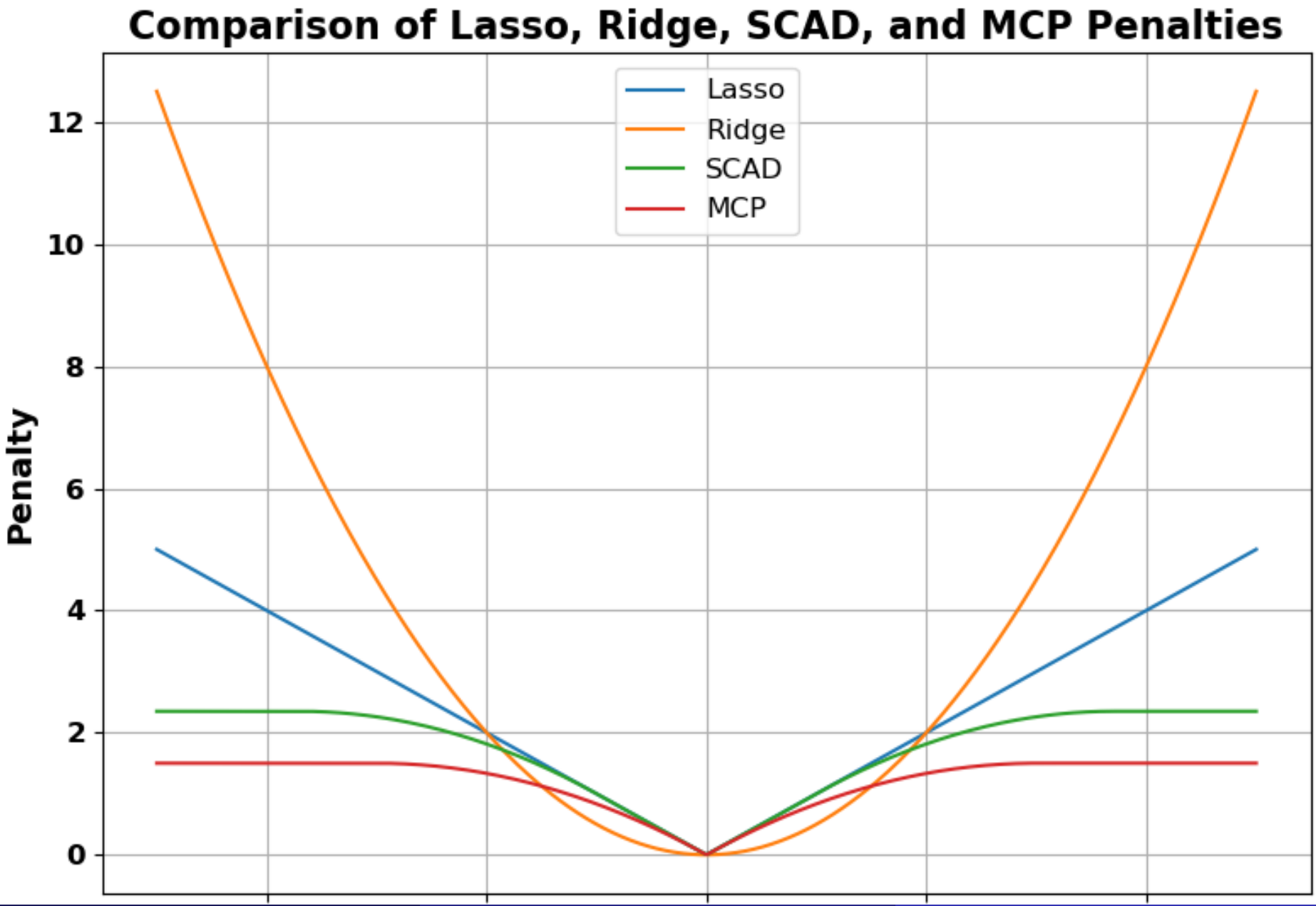
- Similar to SCAD: non-convex, zeros out small coefficients, but spares large ones from heavy penalty.
- Potential **oracle property** with correct tuning.

$$P_\lambda^{\text{MCP}}(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma}, & |\beta| \le \gamma\lambda, \\ \frac{\gamma\lambda^2}{2}, & |\beta| > \gamma\lambda. \end{cases}$$

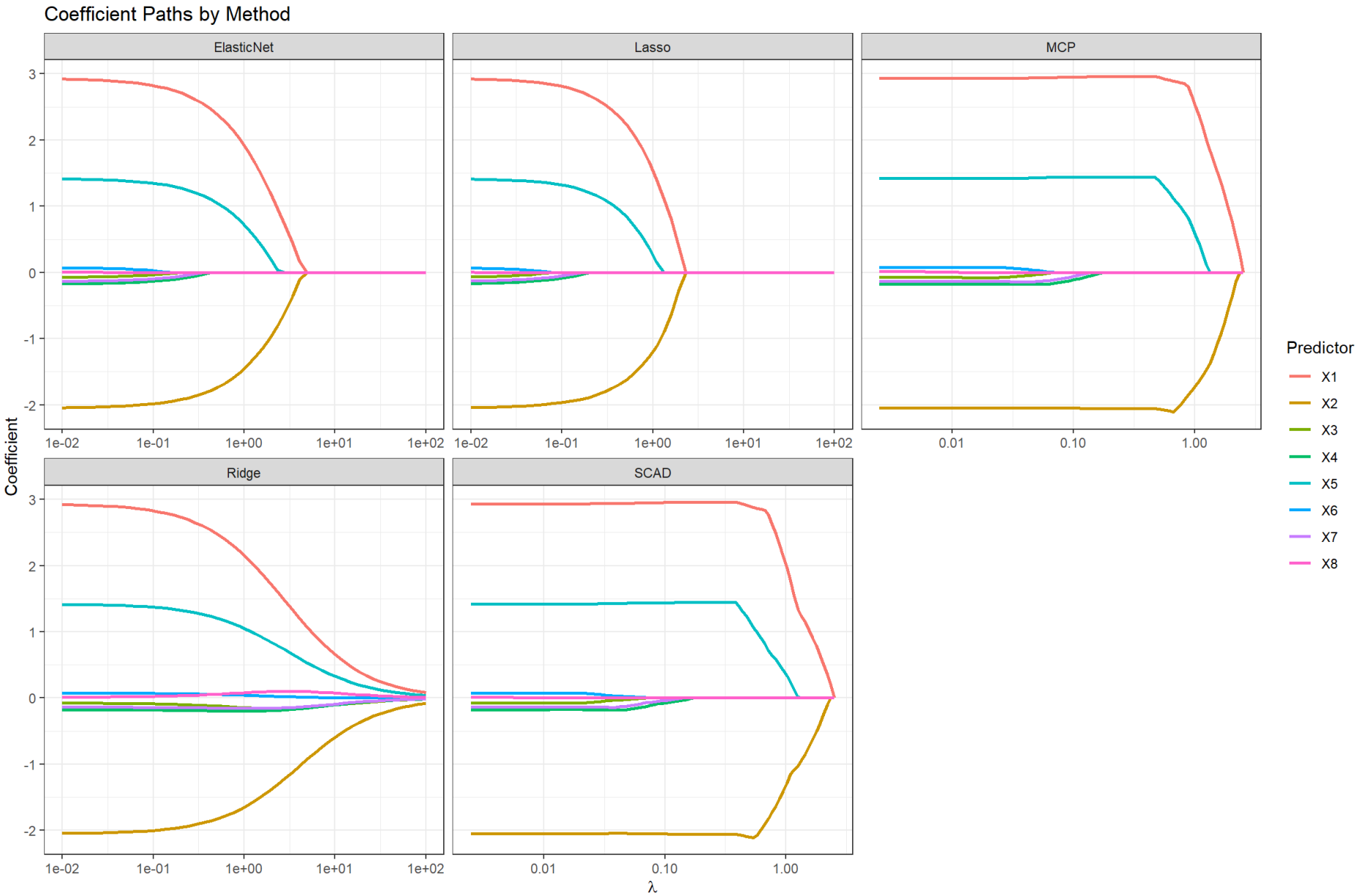
Summary:

- SCAD and MCP differ in details, but both reduce "large-coefficient bias" vs. Lasso.
- Implementable in `ncvreg` (R) for linear/logistic/Cox models.

Penalty Functions by Method



Coefficient Paths by Method



Outline

1. Motivation
2. Five Methods: Ridge, Lasso, EN, SCAD, MCP
3. Group Selection
4. Practical Tips & Oracle Property
5. References and Q&A

Group Selection: Motivation & Definition

Motivation:

- In high-dimensional bioinformatics, features often come in *natural groups*:
  - Gene sets or pathways
  - Multiple SNPs in the same LD block
  - Proteins in the same complex
- When features in a group should be *jointly* selected or excluded, **group selection** methods are preferable to single-feature selection.

Definition:

- Suppose the parameter vector $\boldsymbol{\beta}$ is partitioned into G groups, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_G)$.
- Group selection aims to shrink $\boldsymbol{\beta}_g$ for entire groups to zero, while retaining or refining the groups that are truly relevant.

Group Selection: Problem Formulation

General Formulation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} w_g \|\boldsymbol{\beta}_g\| \right\},$$

- $\ell(\boldsymbol{\beta})$ is the log-likelihood (or a loss function) as before.
- $\boldsymbol{\beta}_g$ is often the ℓ_2-norm (or $\sqrt{\sum_{j \in g} \beta_j^2}$) of the coefficients in group g.
- w_g are optional group weights (e.g., $\sqrt{|g|}$).
- λ controls overall penalty strength.

Interpretation:

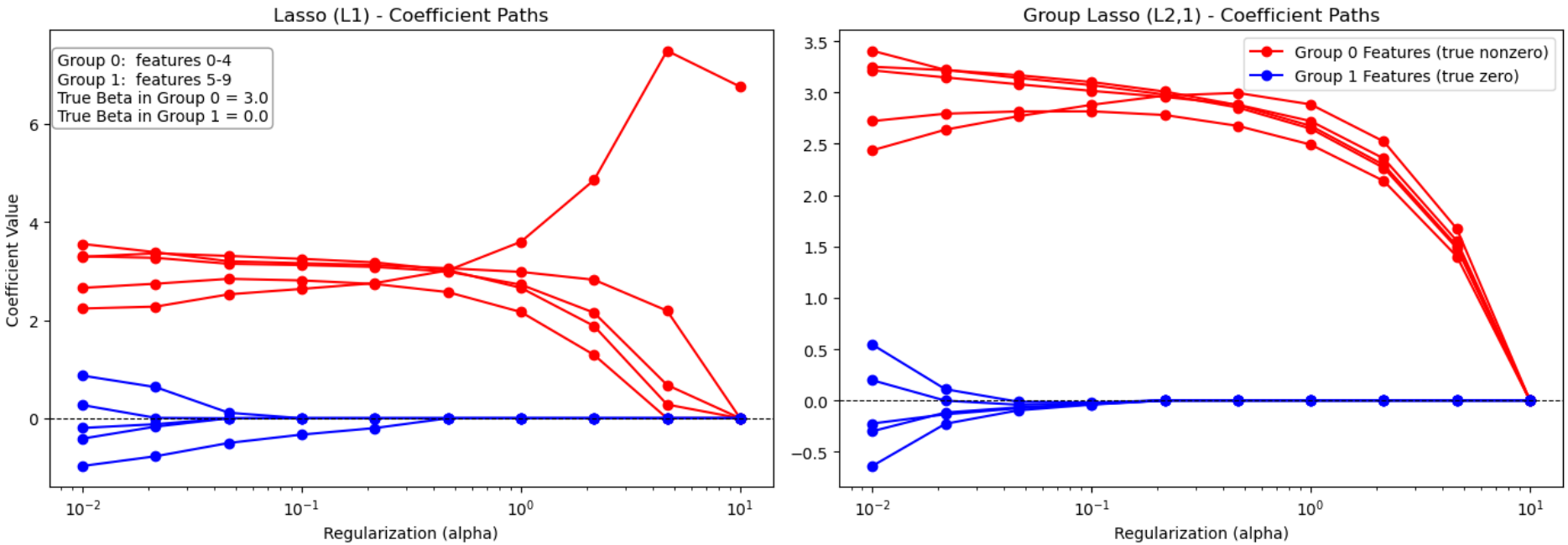
- ℓ_2 penalty on each group encourages *all coefficients* in a group to be zero simultaneously.
- The ℓ_2-norm *within* a group preserves relative weighting, but can zero out the group as a whole if unimportant.

Group Selection

β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	$\left(\beta_{11}^2+\beta_{12}^2+\beta_{13}^2+\beta_{14}^2+\beta_{15}^2+\beta_{16}^2\right)^{1/2}$
β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	$\left(\beta_{21}^2+\beta_{22}^2+\beta_{23}^2+\beta_{24}^2+\beta_{25}^2+\beta_{26}^2\right)^{1/2}$
β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}	$\left(\beta_{31}^2+\beta_{32}^2+\beta_{33}^2+\beta_{34}^2+\beta_{35}^2+\beta_{36}^2\right)^{1/2}$
β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}	$\left(\beta_{41}^2+\beta_{42}^2+\beta_{43}^2+\beta_{44}^2+\beta_{45}^2+\beta_{46}^2\right)^{1/2}$
β_{51}	β_{52}	β_{53}	β_{54}	β_{55}	β_{56}	$\left(\beta_{51}^2+\beta_{52}^2+\beta_{53}^2+\beta_{54}^2+\beta_{55}^2+\beta_{56}^2\right)^{1/2}$
β_{61}	β_{62}	β_{63}	β_{64}	β_{65}	β_{66}	$\left(\beta_{61}^2+\beta_{62}^2+\beta_{63}^2+\beta_{64}^2+\beta_{65}^2+\beta_{66}^2\right)^{1/2}$

$$\Omega(\boldsymbol{\beta}) = \sum_i \sqrt{\sum_j \beta_{ij}^2}$$

Group Selection



Outline

1. Motivation
2. Five Methods: Ridge, Lasso, EN, SCAD, MCP
3. Group Selection
4. Practical Tips & Oracle Property
5. References and Q&A

Cross-Validation & Tuning

Choosing λ:

- K-fold cross-validation is standard, or AIC/BIC/EBIC if feasible.
- For SCAD/MCP, local linear approximation (LLA) or coordinate descent with warm starts is common.
- HPC approaches exist if $p \gg 10^5$ (screening, partial fits).

Common Implementation Tools:

- `glmnet` (R/Python) for Ridge, Lasso, Elastic Net.
- `ncvreg` (R) for SCAD, MCP.

Comparison of Penalties

Penalty	Sparsity?	Bias on Large Coeffs?	Convex?
Ridge (L2)	No	Low	Yes
Lasso (L1)	Yes	High	Yes
Elastic Net	Yes	High	Yes
SCAD	Yes	Lower	No
MCP	Yes	Lower	No

Takeaways:

- SCAD/MCP yield sparser solutions with less bias, but require non-convex optimization.
- Lasso/EN are simpler, widely available, and handle large-scale data easily.

What is the Oracle Property? (1/2)

Definition:

- An estimator has the **oracle property** if, asymptotically:
  1. It correctly identifies zero vs. non-zero coefficients.
  2. It estimates non-zero coefficients as if the true model were already known.

Implication:

- No wasted effort on truly zero variables.
- (Nearly) perfect estimation of large signals with minimal bias.

What is the Oracle Property? (2/2)

Which Penalties Have It?

- **SCAD, MCP** can approach near-oracle behavior under certain conditions.
- **Lasso** generally does not (over-shrinks large coefficients).

Conditions for Oracle:

- $\sqrt{n}$-consistency in coefficient estimates.
- Proper tuning of λ, a (SCAD), γ (MCP).
- Adequate separation between zero and non-zero signals.

Practical Reminders

- **Cross-validation** is crucial for choosing λ.
  - Repeated CV or stability selection if n is small.
- **Correlation Check**:
  - Lasso might select only one from a correlated block; EN or Ridge can share weights.
- **Validation**:
  - External cohorts or hold-out sets if possible.
- **Interpretability vs. Performance**:
  - Ridge retains all features, safer if all may matter.
  - Lasso/SCAD/MCP/EN yield sparser, more interpretable solutions.

Outline

1. Motivation
2. Five Methods: Ridge, Lasso, EN, SCAD, MCP
3. Group Selection
4. Practical Tips & Oracle Property
5. References and Q&A

Q & A

Questions?

References (1/2)

Hoerl, A. E. & Kennard, R. W. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics*, **12**(1), 55–67.

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B*, **58**(1), 267–288.

Zou, H. & Hastie, T. (2005). *Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B*, **67**(2), 301–320.

Fan, J. & Li, R. (2001). *Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association*, **96**(456), 1348–1360.

References (2/2)

- **Zhang, C.-H.** (2010). *Nearly Unbiased Variable Selection under Minimax Concave Penalty. Annals of Statistics*, **38**(2), 894–942.
- **Takeuchi, F.** & **Kato, N.** (2021). *Nonlinear Ridge Regression Improves Cell-Type-Specific Differential Expression Analysis. BMC Bioinformatics*, **22**, 355.
- **Das, J.**, **Gayvert, K. M.**, **Bunea, F.**, **Wegkamp, M. H.**, & **Yu, H.** (2015). *ENCAPP: Elastic-net-based Prognosis Prediction and Biomarker Discovery for Human Cancers. BMC Genomics*, **16**, 1–13.
- **Jiang, Y.**, **Montierth, M. D.**, **Yu, K.**, **Ji, S.**, **Guo, S.**, **Tran, Q.**, **Liu, X.**, **Shin, S. J.**, **Cao, S.**, **Li, R.**, *et al.* (2024). *Pan-cancer Subclonal Mutation Analysis of 7,827 Tumors Predicts Clinical Outcome. bioRxiv*, 2024–07.
- **Lim, D. K.**, **Rashid, N. U.**, & **Ibrahim, J. G.** (2021). *Model-Based Feature Selection and Clustering of RNA-Seq Data for Unsupervised Subtype Discovery (FSCseq). The Annals of Applied Statistics*, **15**(1), 481–508.