Link Violation in GLMs

Helian Feng, Linglin Huang, Andy Shi

BST 235 Final Project

December 11, 2017

Introduction

- **Question of Interest:** Recover the coefficients and get reasonable prediction under a misspecified model (link violation).
- **Setting:** iid $\{(\mathbf{X}_i, Y_i), i = 1, ..., n\}$.
- Y_i: continuous response.
- $\mathbf{X}_i$: $p \times 1$ vector of covariates.
- True model

$$Y_i = \alpha_0 + \Phi(\boldsymbol{\beta}_\mathbf{0}^\mathbf{T}\mathbf{X_i}) + \epsilon_i, \quad \epsilon_i \perp X_i, \text{ mean 0, variance } \sigma^2.$$

- However, we fit the linear model

$$Y_i = \alpha + \boldsymbol{\beta}^\mathbf{T}\mathbf{X_i} + \epsilon_i$$

with adaptive LASSO, get estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

Goals

- If $\mathbf{X}_i$ follows a multivariate normal distribution, estimates from adaptive LASSO under misspecified link ($\hat{\boldsymbol{\beta}}$) converges to a limit ($\bar{\boldsymbol{\beta}}$), which is proportional to the true coefficients ($\boldsymbol{\beta_0}$). Thus, we can consistently estimate the support of $\boldsymbol{\beta_0}$.
- Perform simulation studies to support.
- Investigate violations of MVN assumption, when $\mathbf{X_i}$ is generated from some non-elliptical distribution.
- Compare prediction performances of:
  - true model
  - adaptive LASSO under link violation
  - adaptive LASSO under link violation with nonparametric calibration
- Nonparametric calibration for a new observation $\mathbf{x}^0$:

$$\hat{m}(\mathbf{x}^0) = \frac{\sum_{i=1}^{n} K_h(\hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X_i} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{x^0})Y_i}{\sum_{i=1}^{n} K_h(\hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X_i} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{x^0})}$$

Theoretical Results

If $\mathbf{X_i}$ follows a multivariate normal distribution, $\hat{\boldsymbol{\beta}}$ converges to a limit, $\bar{\boldsymbol{\beta}}$, that proportional to $\boldsymbol{\beta_0}$. Thus it can also consistently estimate the support of $\boldsymbol{\beta_0}$.

Proof overview:

1. $\hat{\boldsymbol{\beta}}$ converges to $\bar{\boldsymbol{\beta}}$ by consistency of adaptive LASSO in Zou (2006).
2. $\bar{\boldsymbol{\beta}}$ is proportional to $\boldsymbol{\beta_0}$ by Theorem 2.2 in Li & Duan (1989).

Proof

Li & Duan (1989): Define

$$R(\alpha, \boldsymbol{\beta}) = \mathbb{E}L(\alpha + \boldsymbol{\beta}\mathbf{x}, y),$$

where $L(\theta, y)$ is any minimization criterion that is convex in θ, and

$$\Omega = \{(\alpha, \boldsymbol{\beta}) : R(\alpha, \boldsymbol{\beta}) \text{ is well-defined and is finite}\}.$$

Hence fitting the regression model can be represented as:

$$\textit{minimize} \quad R(\alpha, \boldsymbol{\beta}) \quad \text{over} \quad (\alpha, \boldsymbol{\beta}) \in \Omega$$

Proof cont.

Theorem 2.2 For a model of the form

$$y = g(\alpha + \boldsymbol{\beta}\mathbf{x}, \epsilon), \epsilon \sim F(\epsilon),$$

where $g(\cdot, \cdot)$ is a given link function, if

- $\mathbf{x_i}$ is normally distributed;
- Ω is a nonempty convex set in $\mathbb{R}^{p+1}$;
- $(\alpha^*, \boldsymbol{\beta}^*)$ the minimizer of R over Ω;

then $\boldsymbol{\beta}^*$ is proportional to $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^* = \gamma\boldsymbol{\beta}$$

for some scalar γ.

Proof cont.

Back to the proof:

- By Zou (2006), $\hat{\boldsymbol{\beta}}^{\text{ALASSO}} \to \bar{\boldsymbol{\beta}}$, where $(\bar{\alpha}, \bar{\boldsymbol{\beta}})$ minimizes $R(\alpha, \boldsymbol{\beta}) = \mathbb{E}(y - \alpha - \boldsymbol{\beta}\mathbf{x})^2$.
- Ω is a nonempty convex set in $\mathbb{R}^{p+1}$
- $\mathbf{X_i}$ follows a multivariate normal distribution;
- $(\bar{\alpha}, \bar{\boldsymbol{\beta}})$ minimizer of R over Ω ;

Therefore $\bar{\boldsymbol{\beta}} = \gamma\boldsymbol{\beta_0}$ for some scalar γ.
So,

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \bar{\boldsymbol{\beta}} = \gamma\boldsymbol{\beta_0}$$

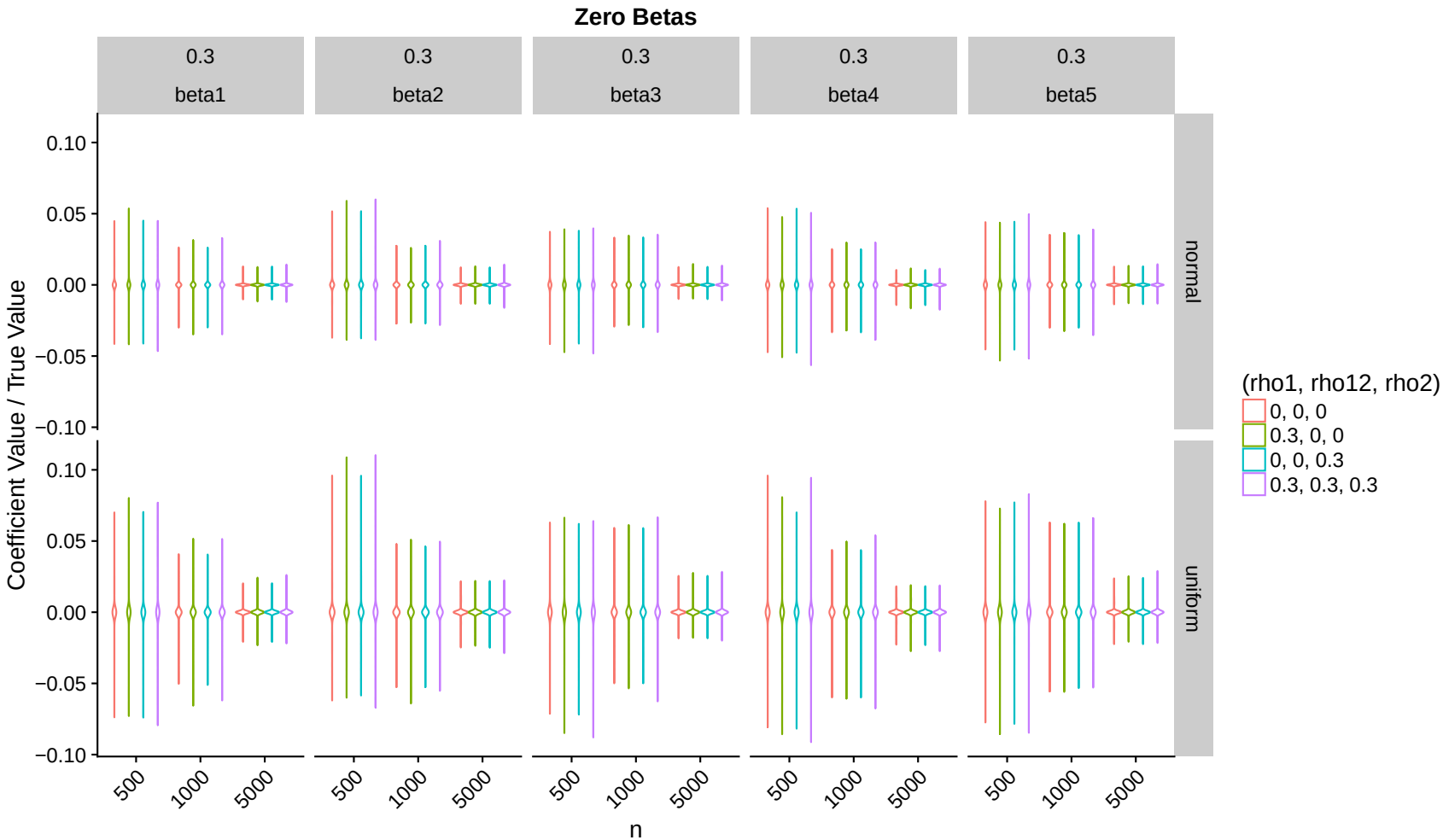
Simulation

- $Y_i = \alpha_0 + \Phi(\boldsymbol{\beta_0^T X_i}) + \epsilon_i, \quad \epsilon_i \perp \mathbf{X_i}$
- Number of iterations: 1000
- Training sample size: 500, 1000, 5000
- Testing sample size: 10000
- $\alpha_0 = 3$, $\boldsymbol{\beta_0} = (0, 0, 0, 0, 0, -0.3, -0.1, 0.1, 0.3)^T$
- $\epsilon_i \sim N(0, \sigma^2)$, $\sigma \in \{0.01, 0.3\}$ (small and big errors)
- $\mathbf{x_i} \sim N(\mathbf{0}, \Sigma)$ (elliptical) or $\mathbf{x_i} \sim \text{Unif}(-1, 1, \Sigma)$ (non-elliptical), where for $\rho_1, \rho_2, \rho_{12} \in \{0, 0.3\}$, Σ is in the form:

$$\Sigma = \begin{array}{c} \\ \text{Zero betas} \\ \text{Nonzero betas} \end{array} \begin{array}{c} \begin{array}{cc} \text{Zero betas} & \text{Non-zero betas} \end{array} \\ \begin{bmatrix} \boxed{\rho_1} & \boxed{\rho_{12}} \\ \boxed{\rho_{12}} & \boxed{\rho_2} \end{bmatrix} \end{array}$$

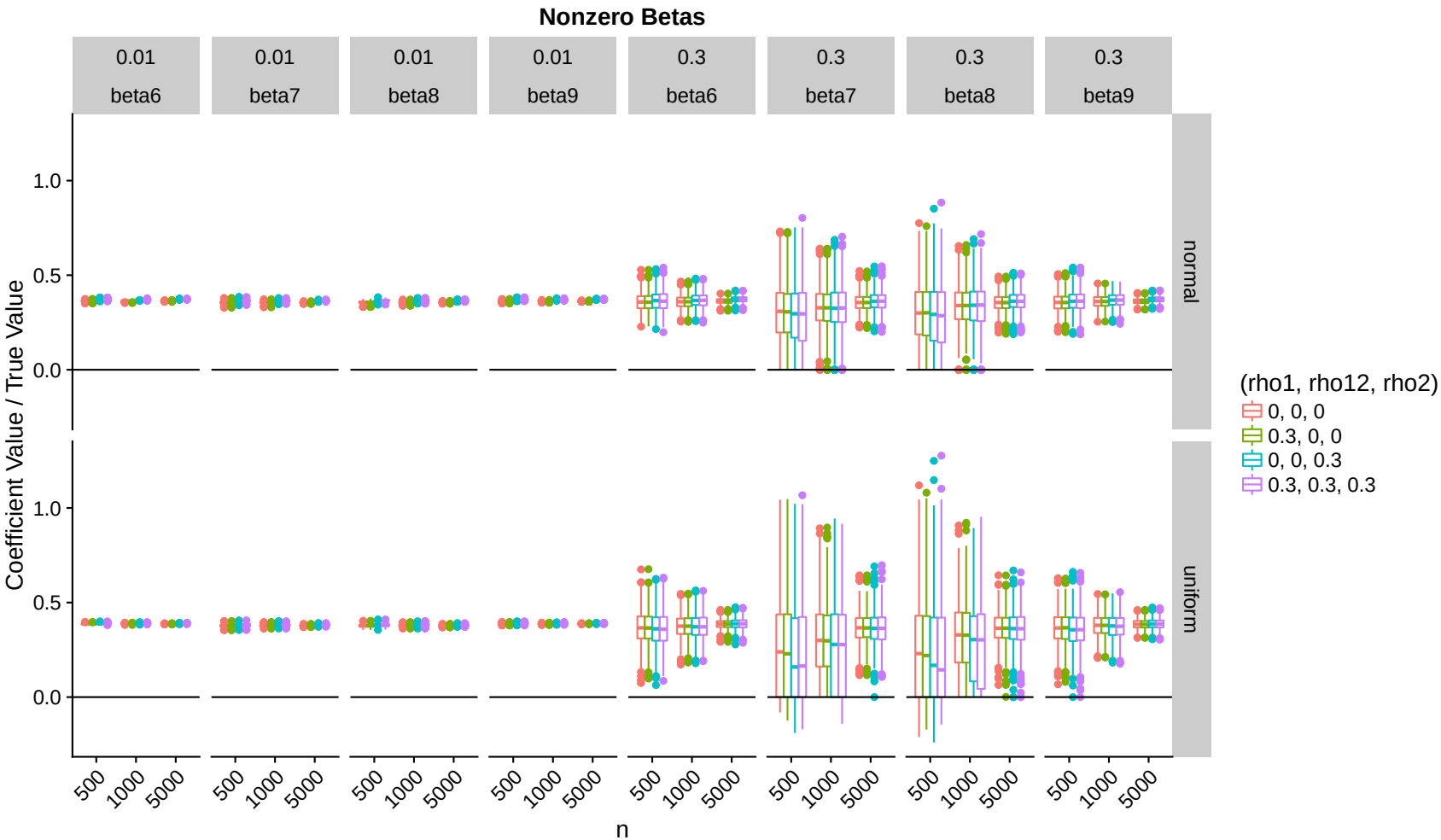
Simulation: Zero β_0

- For small errors, all true zero $\beta = 0$.
- Plot for big errors shown here:



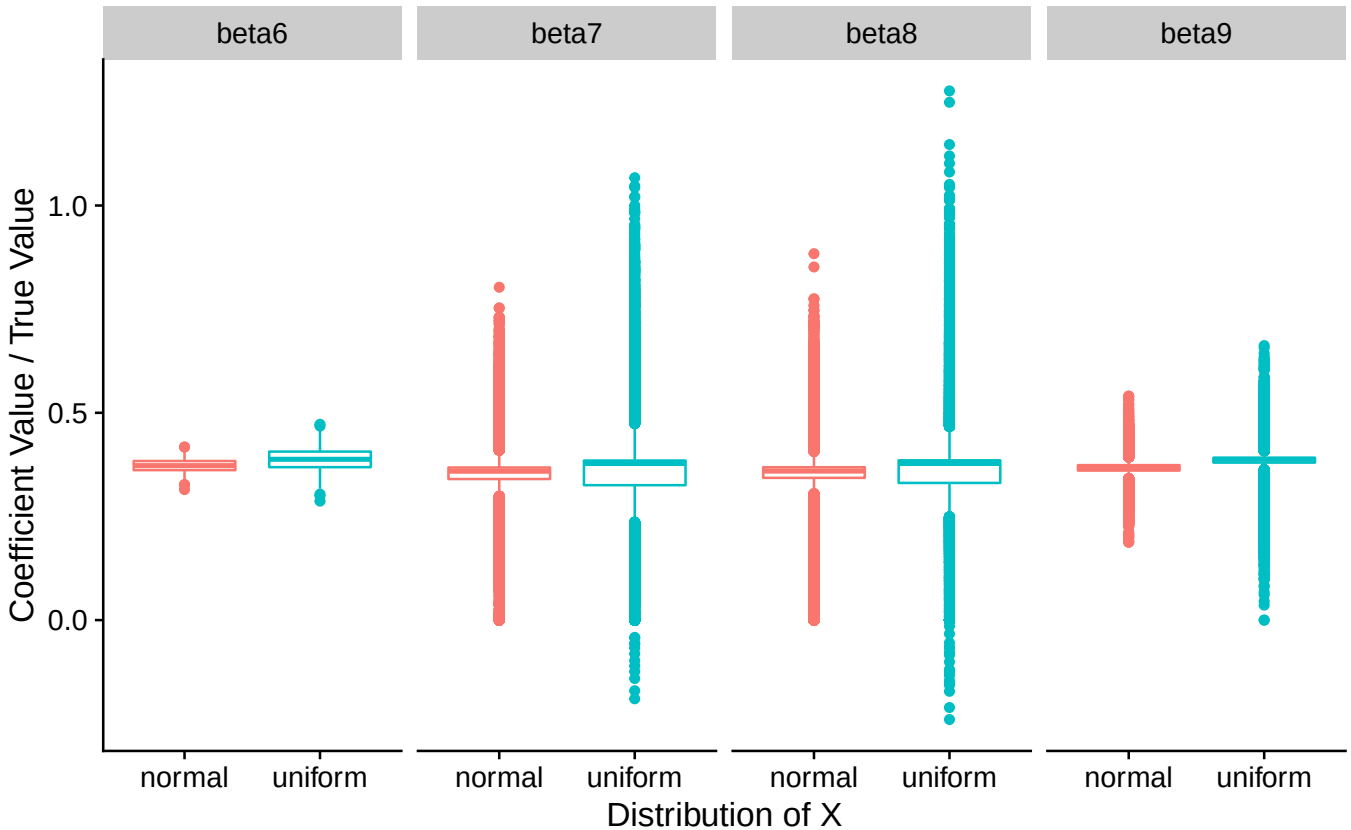
Simulation: Nonzero β_0

- Can recover non-zero β up to a scalar.



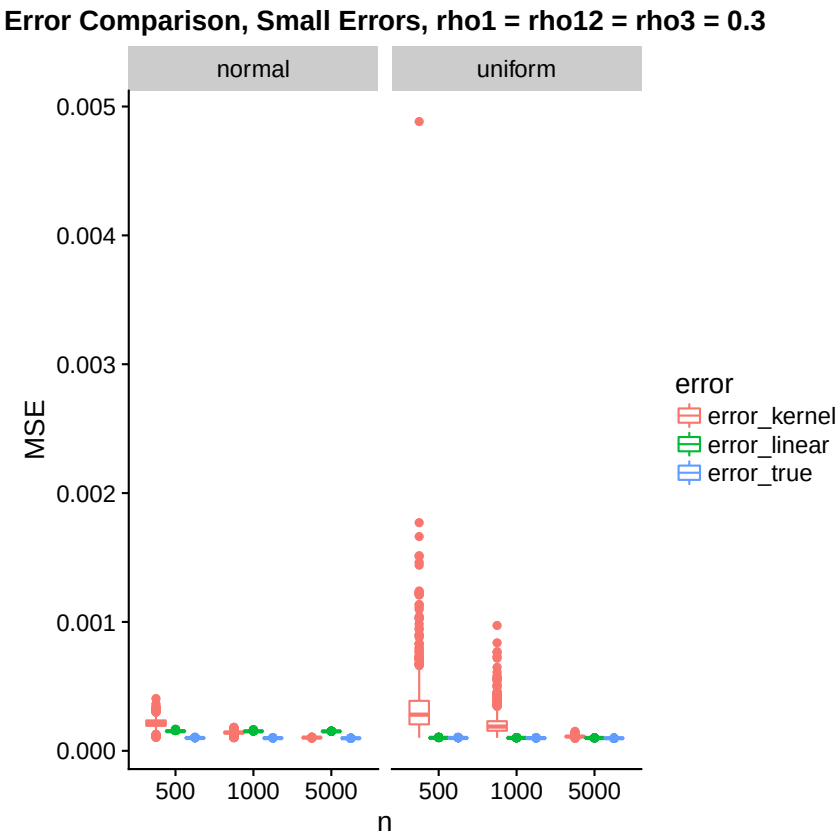
Simulation: Nonzero β_0 Proportionality

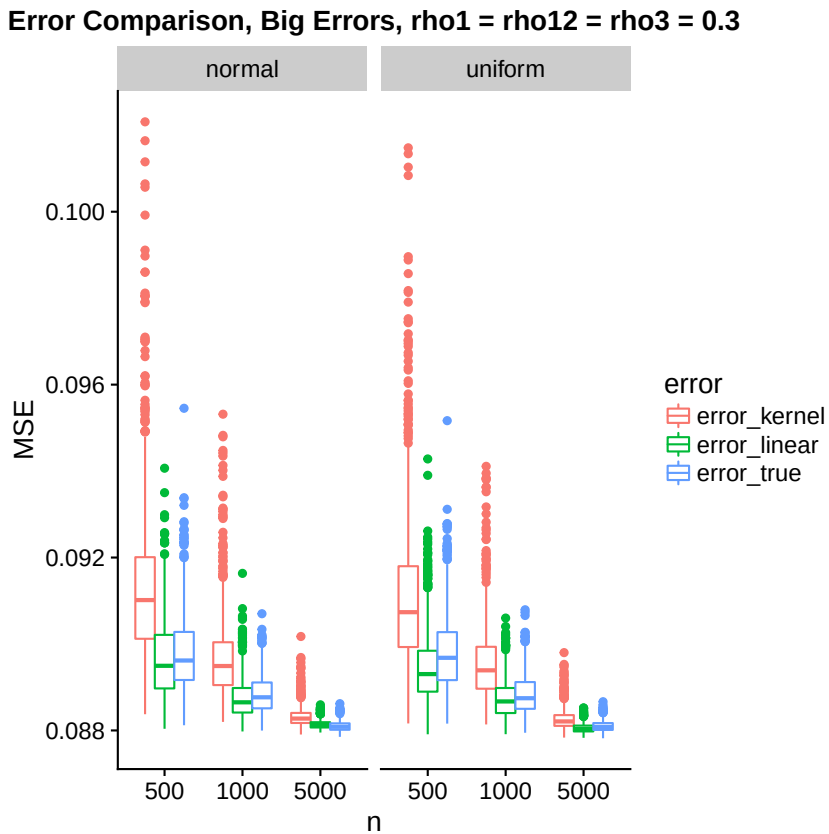
- When $\mathbf{X}_i \sim$ Uniform, $\hat{\boldsymbol{\beta}}$ has bigger variance, but still proportional to the true $\boldsymbol{\beta_0}$.
- Plot shown for $n = 5000, \rho_1 = \rho_{12} = \rho_2 = 0.3$, big errors:



Simulation: Prediction Performance

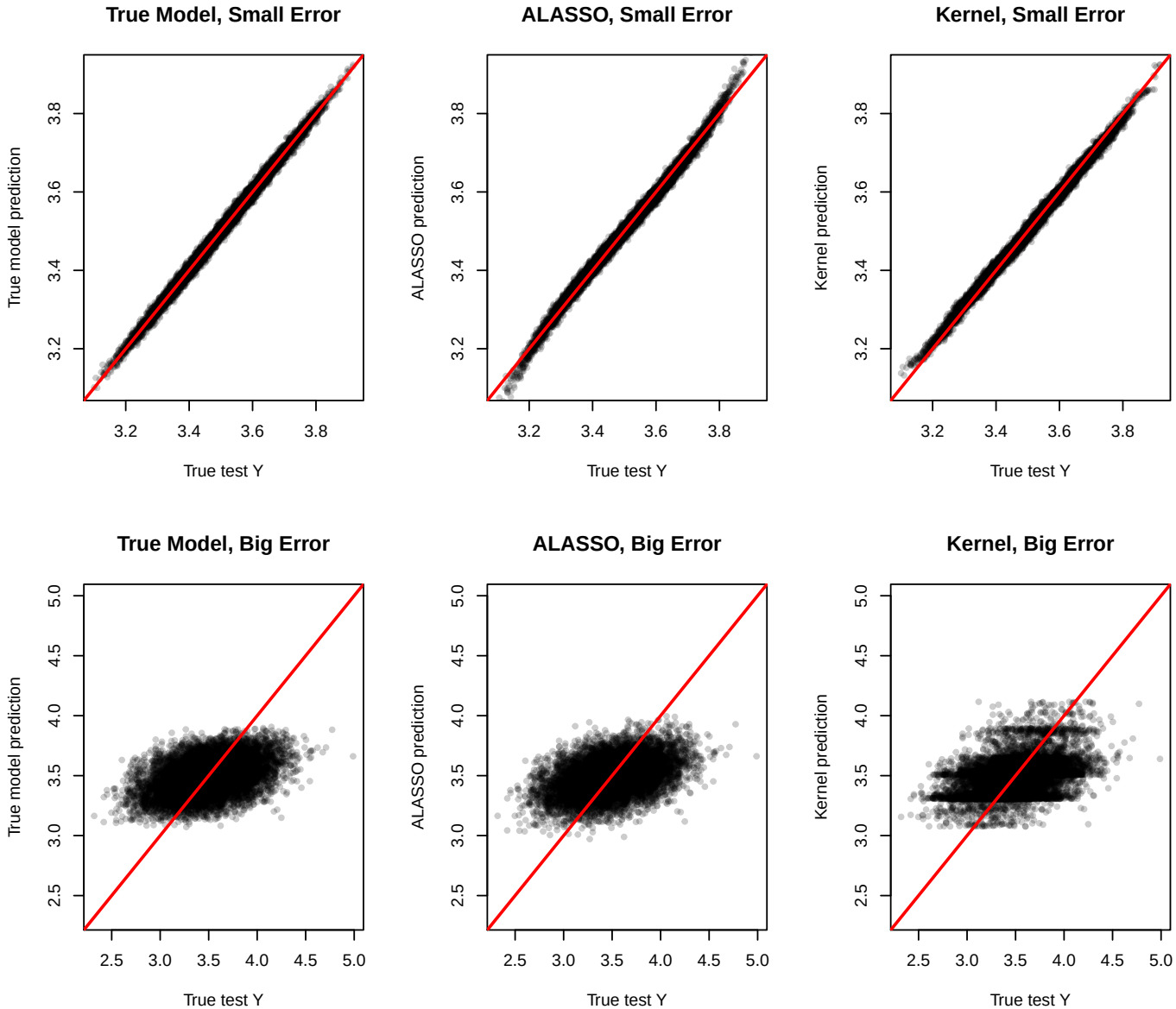
- Very similar for different correlation structures for predictors.
- Shown here: MSEs for $\rho_1 = \rho_{12} = \rho_2 = 0.3$.
- True model and kernel smoothing seem to perform badly, especially with bigger noise.





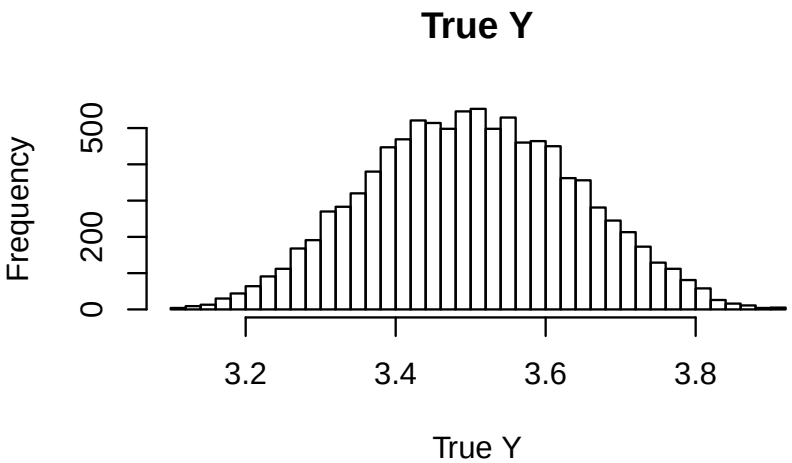
Prediction Error Exploration: Scatterplots

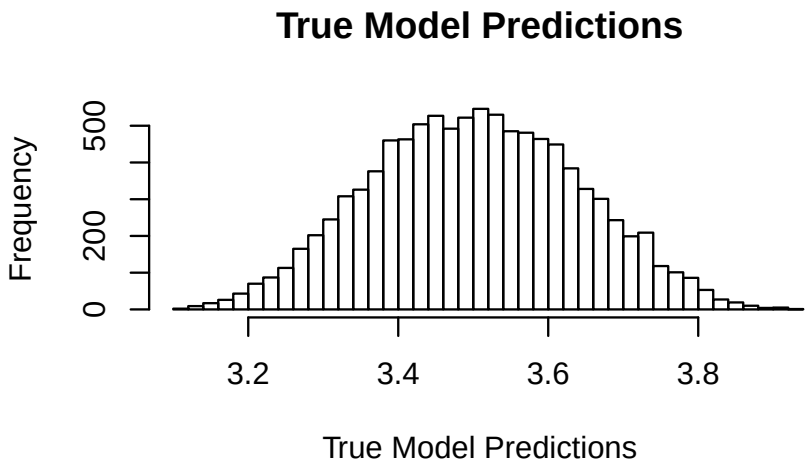
Relationship between predictions and true Y:

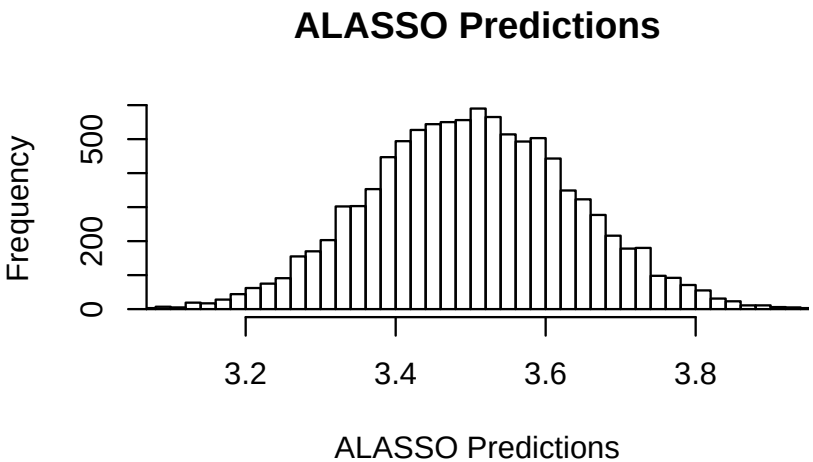


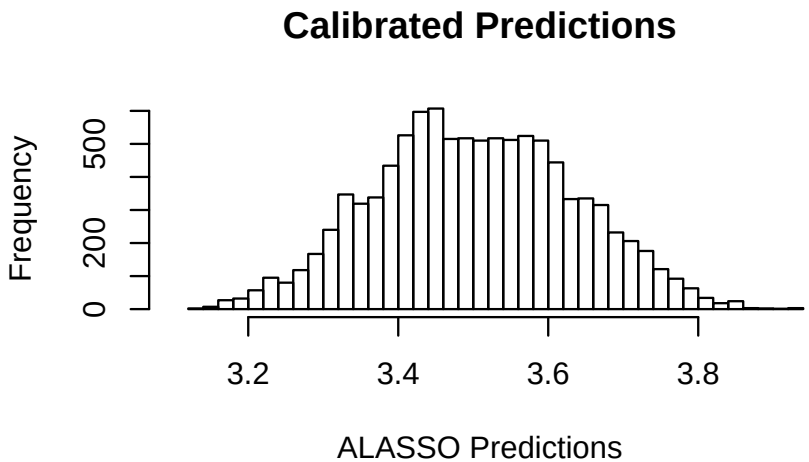
Prediction Error Exploration: Small Error Histograms

- Histogram of predictions with small error.
- Predictions DO cover the range of true Y.



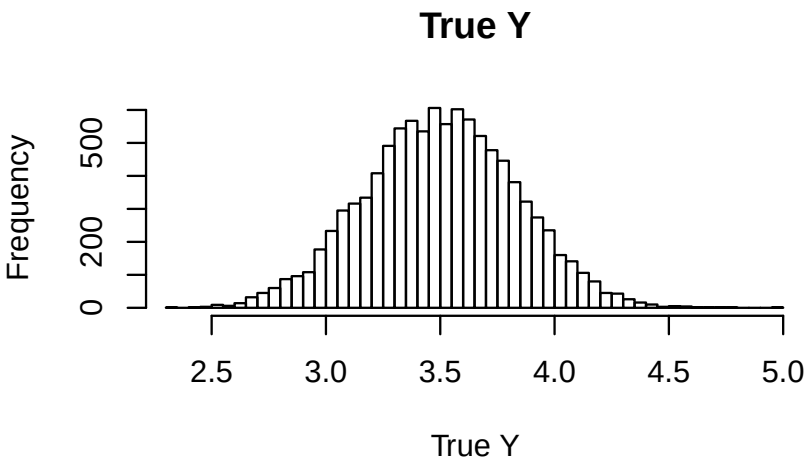


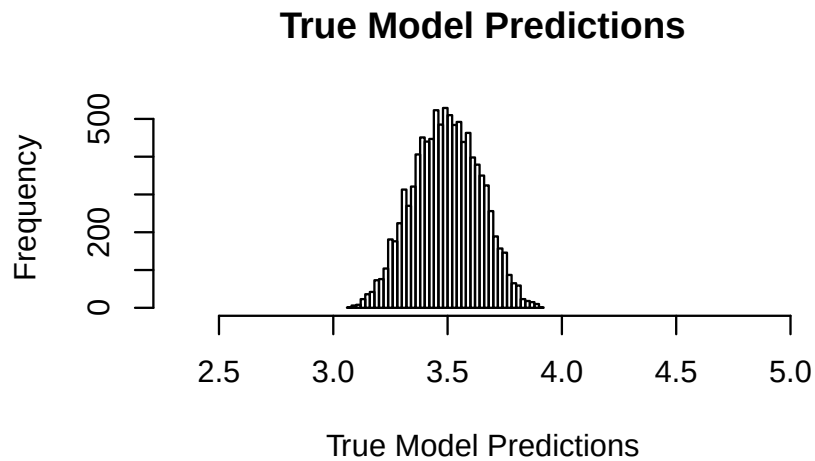


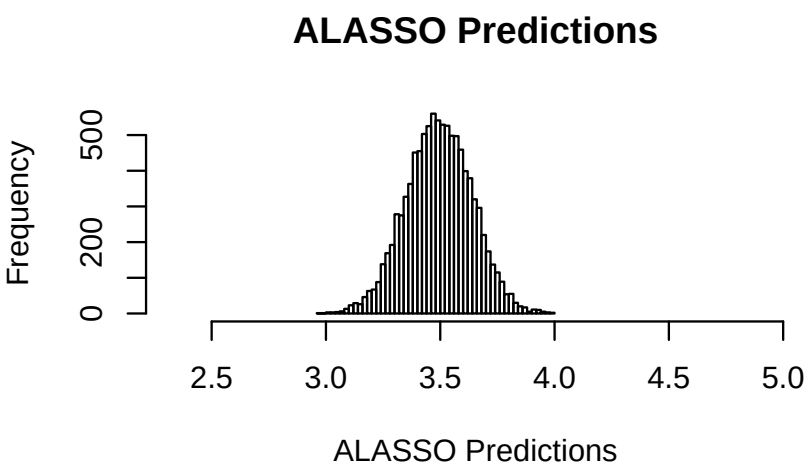


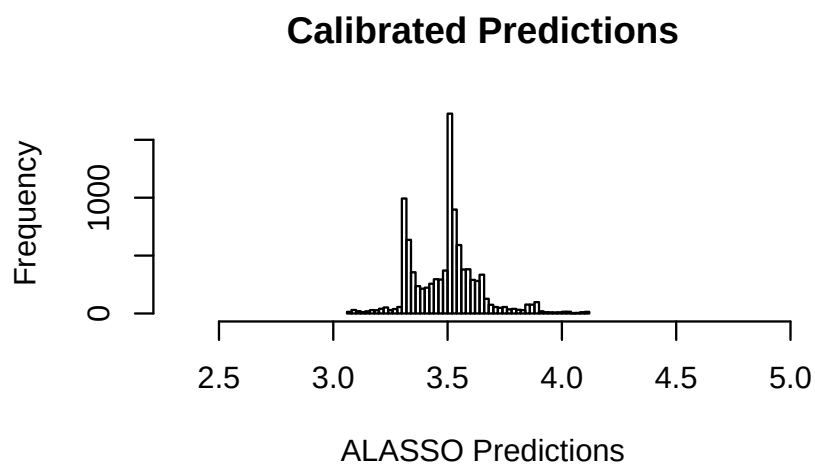
Prediction Error Exploration: Big Error Histograms

- Histogram of predictions with large error.
- Predictions DO NOT cover the range of true Y.



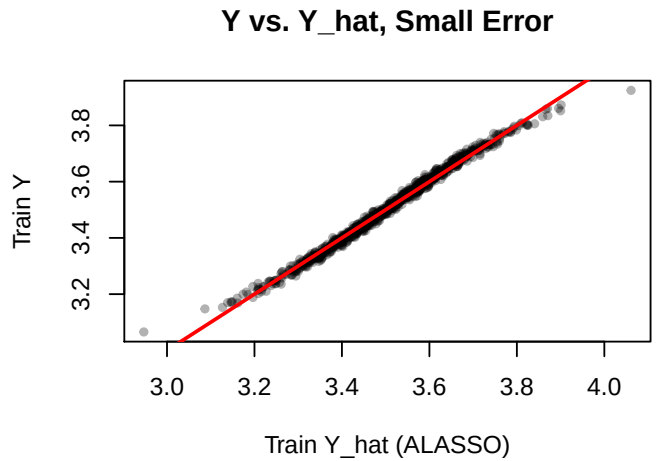


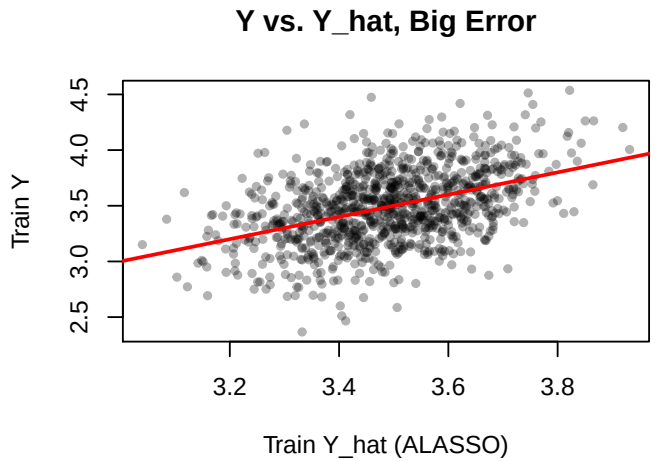


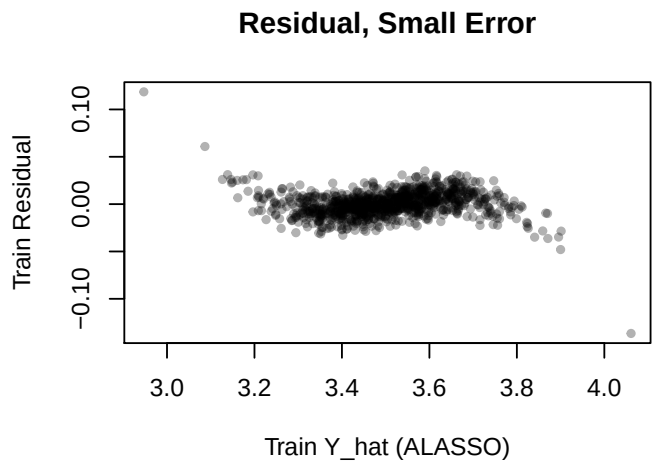


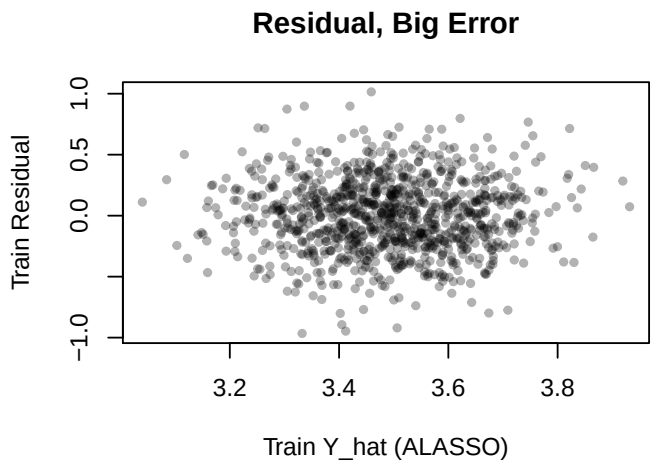
Prediction Error Exploration: Residuals

- Small errors: Residuals have a pattern that can be recovered using kernel smoothing.
- Large errors: Residuals look random.





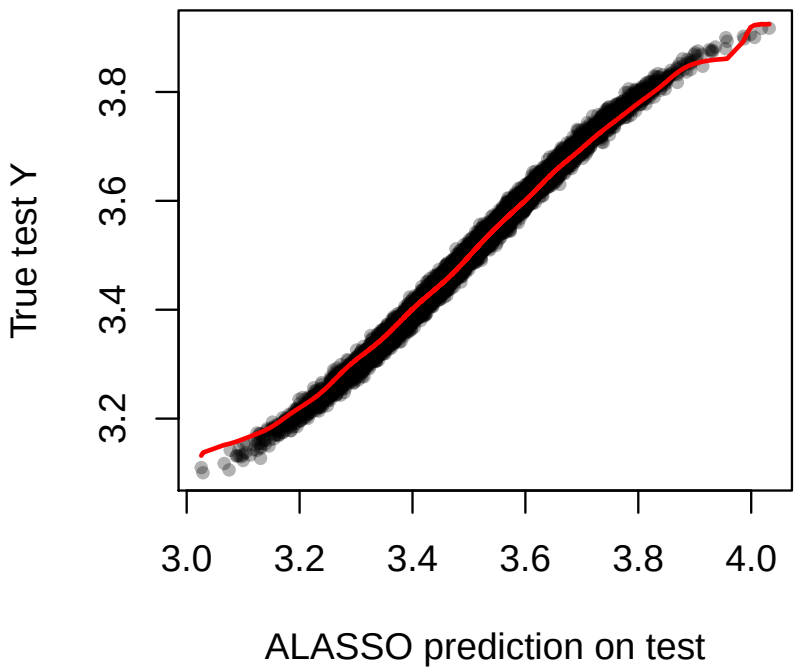




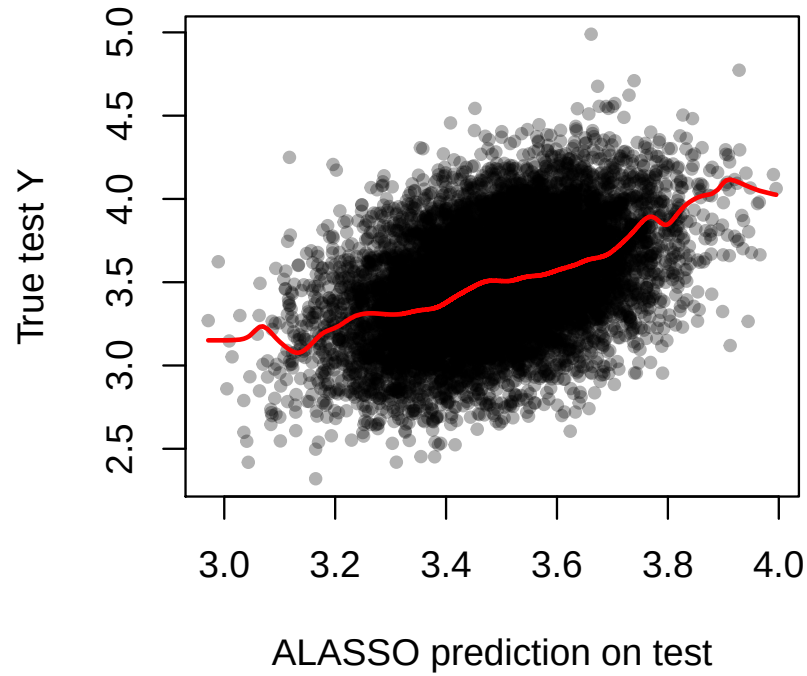
Prediction Error Exploration: Kernel Smoothing

- Small errors: Kernel can do smoothing to improve predictions.
- Large errors: Kernel smoothing does not help because no clear pattern between adaptive LASSO predictions and true Y.

Kernel Predictions, Small Error



Kernel Predictions, Big Error



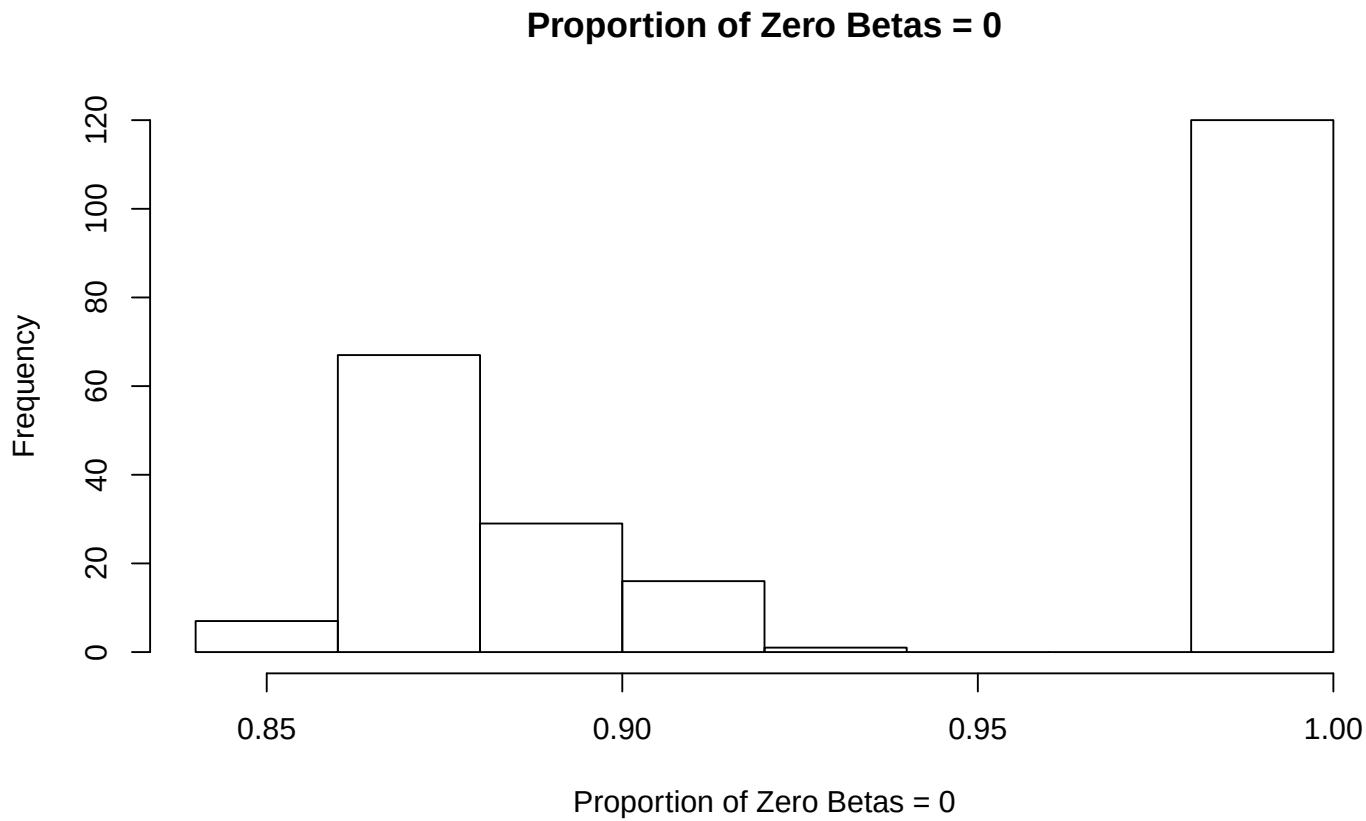
Discussion

- In theory, adaptive LASSO can recover the true coefficients proportionally, even under link violation.
- In simulation, the model successfully estimates the support of true coefficients.
- Seems robust when the elliptical distribution assumption is violated.
- When the noise is small, adaptive LASSO has bigger MSE than the true model, and smoothing over the adaptive LASSO fit helps lower the MSE.
- When the noise is big, the predictions are poor for all three models.
  - True model suffers from constrained prediction space.
  - Adaptive LASSO works slightly better.
  - Kernel smoothing on top of adaptive LASSO fit doesn't help. The non-linear relationship between the outcome and adaptive LASSO fit is masked by the noise.
- The correlation structure between the predictors does not seem to matter much.

Thank you

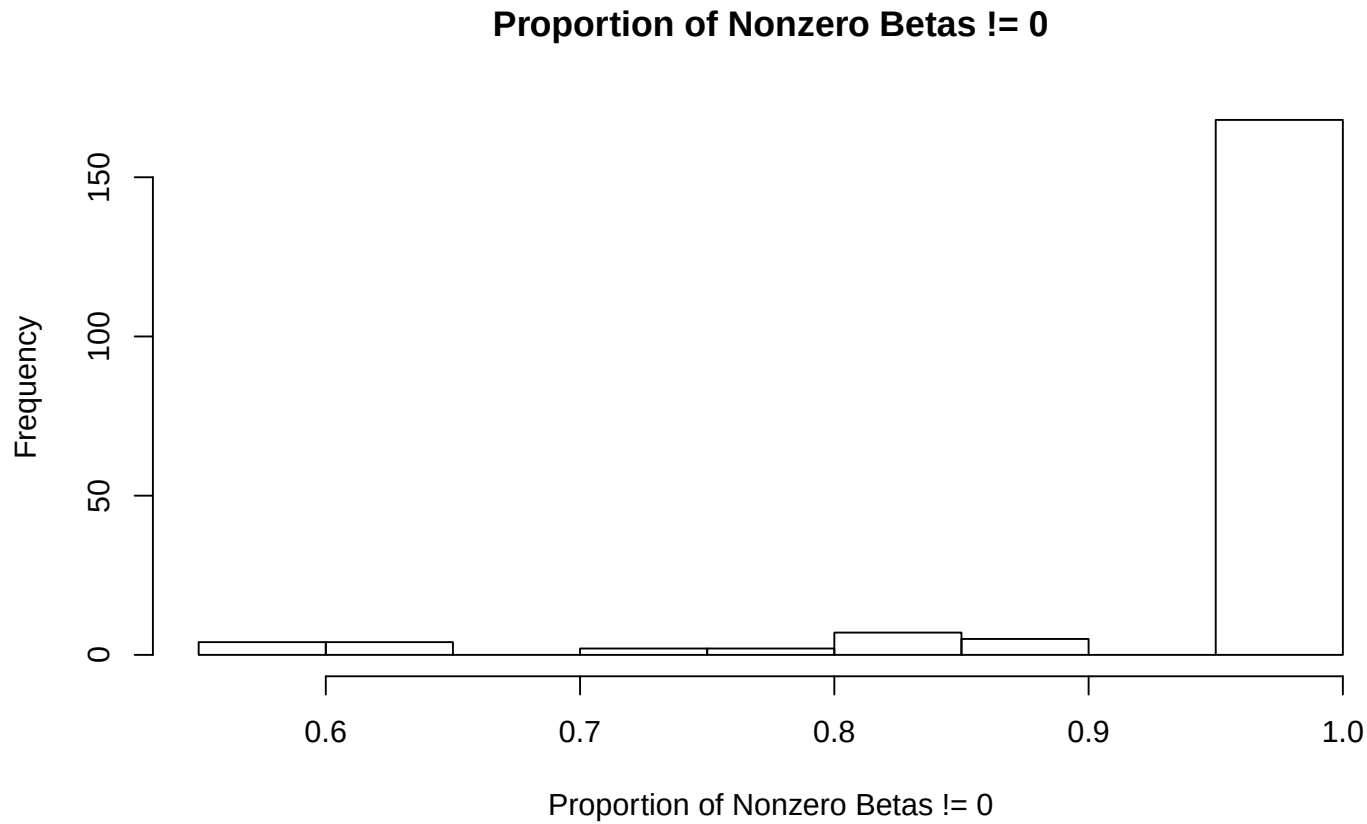
Appendix: Proportion Zero

Proportion of simulated $\hat{\beta}_j = 0$ when true $\beta_{0j} = 0$, over all simulation conditions.

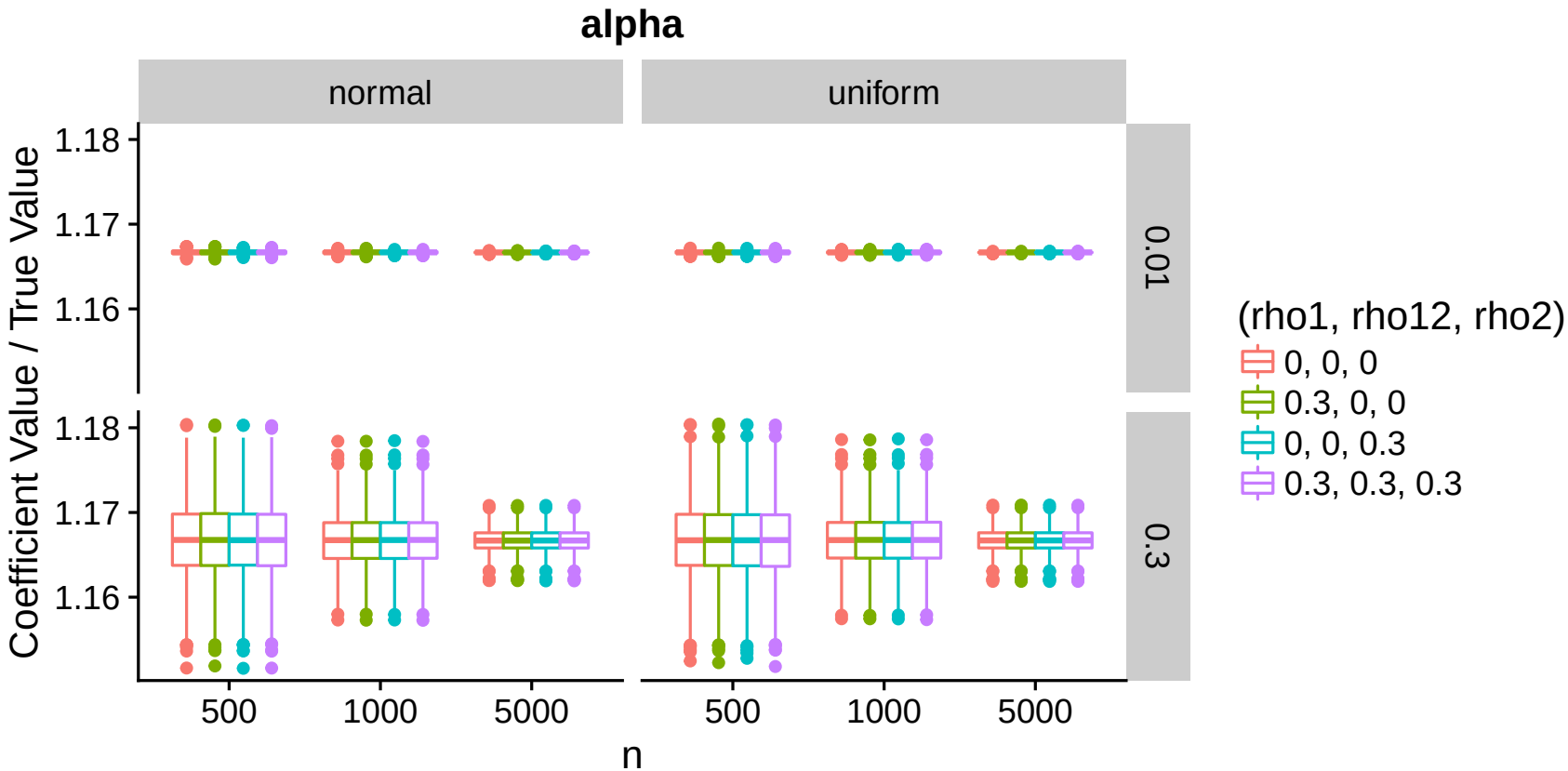


Appendix: Proportion Nonzero

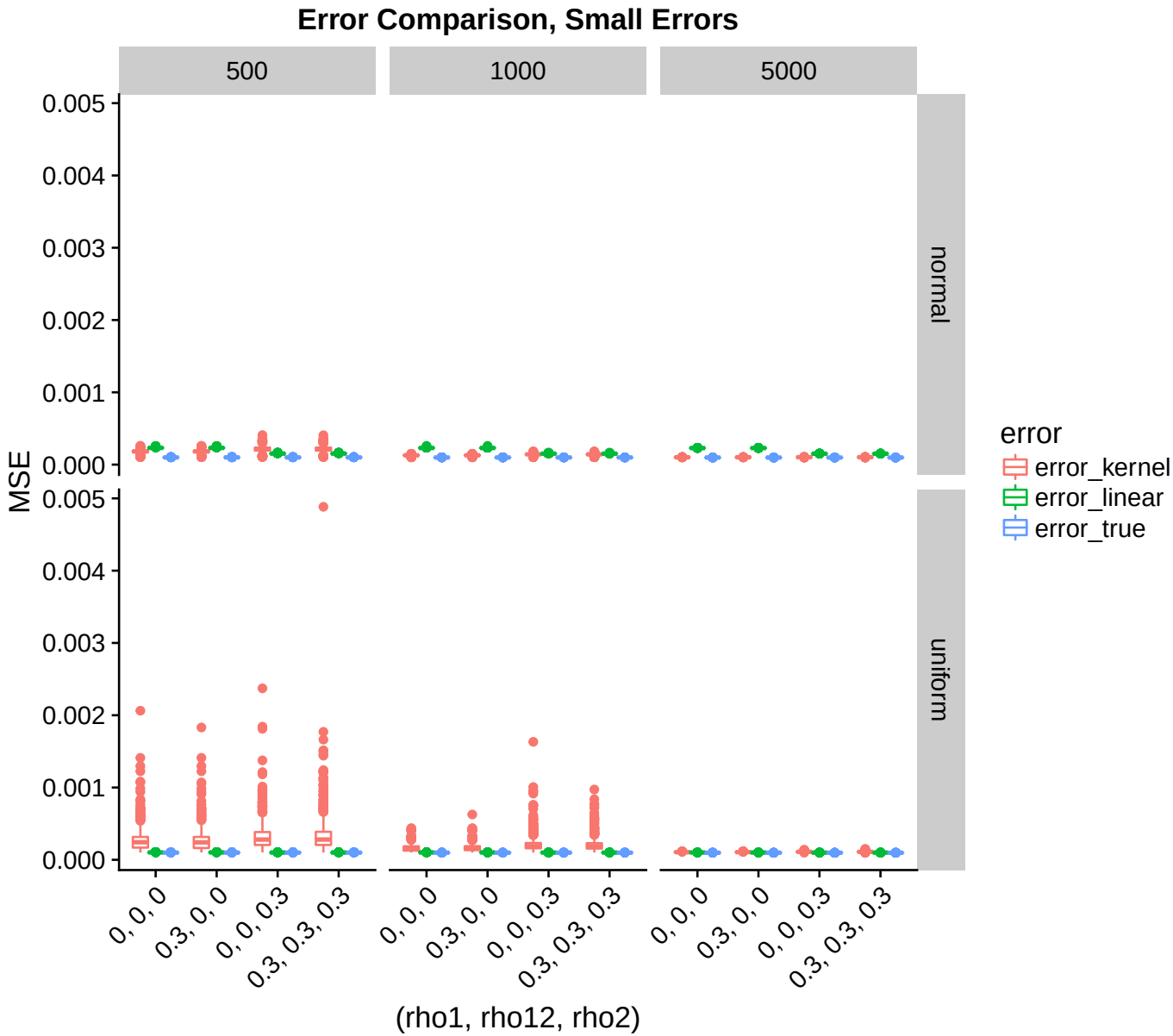
Proportion of simulated $\hat{\beta}_j \neq 0$ when true $\beta_{0j} \neq 0$, over all simulation conditions.



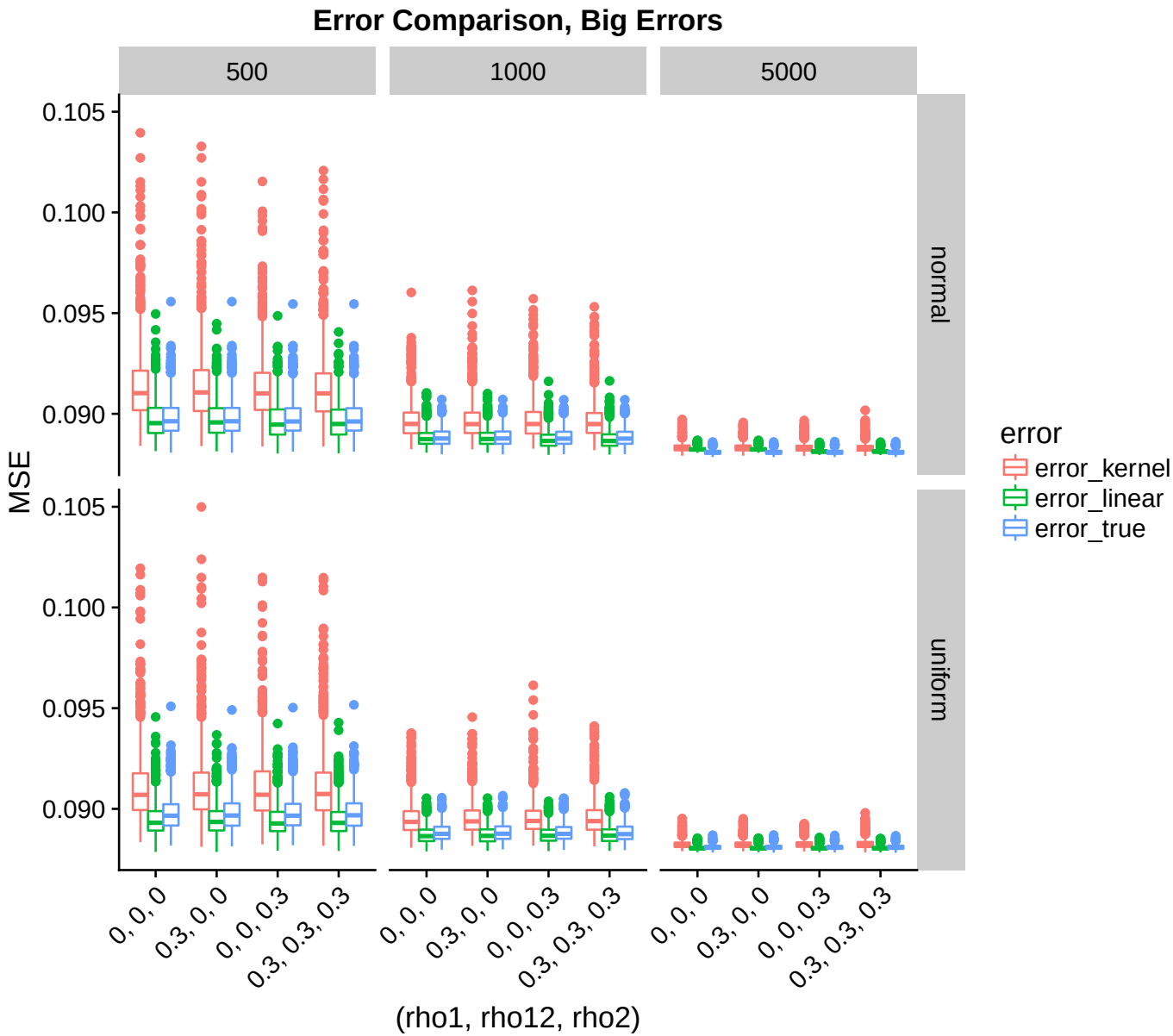
Appendix: α



Appendix: Prediction Performance Full Plots



Appendix: Prediction Performance Full Plots



Appendix: Elliptical Distributions

A random vector X on a Euclidean space has an **elliptical distribution** if its characteristic function ϕ satisfies the following functional equation (for every column-vector t)

$$\phi_{X-\mu}(t) = \phi(t^T \Sigma t)$$

for some location parameter μ and some nonnegative-definite matrix Σ.

Examples:

- Multivariate normal distribution
- Multivariate t-distribution
- Symmetric multivariate stable distribution
- Symmetric multivariate Laplace distribution
- Multivariate logistic distribution
- Multivariate symmetric general hyperbolic distribution

Appendix: Generating Correlated Uniforms

Using the copula method:

1. Generate $(X_{i1}, \ldots, X_{ip}) \sim N(0, \Sigma)$ where Σ is a *correlation* matrix (1s on diagonal).
2. Then, get $(U_{i1}, \ldots, U_{ip}) = (\Phi(X_{i1}), \ldots, \Phi(X_{1p}))$. Marginally, $U_{ij} \sim \text{Uniform}(0, 1)$.
3. Transform $(b - a)U_{ij} + a$ to get $U_{ij} \sim \text{Uniform}(a, b)$.