1. (a) Suppose that $U$ and $V$ are bivariate Gaussian, and $g$ is a function. Show that

$$ \text{cov}(g(U), V) = \text{cov}(U, V) \text{cov}(g(U), U) / \text{var}(U). $$

(b) Suppose that we have a “single index model” in which

$$ Y = g(\alpha + \beta'X) + \epsilon, $$

Based on part (a), show that if $X$ is multivariate Gaussian with a non-singular covariance matrix, then OLS regression of $Y$ on $X$ yields $\hat{Y} = \hat{\alpha} + X \hat{\beta}$, where $\hat{\beta}$ consistently estimates $k\beta$ for a constant $k \in \mathbb{R}$.

(c) Use linearization to derive an approximate standard deviation for $\hat{\beta}_j/\hat{\beta}_k$ in the single index model.

(d) Use simulation to assess the estimates and standard errors when $g(z) = \text{arctan}(z)$.

2. Suppose we have two matrices $X_a$ and $X_b$, where $X_a$ is $n \times p + 1$, $X_b$ is $n \times p + q$, and $\text{col}(X_a) \subset \text{col}(X_b)$. We observe $Y = EY + \epsilon$, where $EY \in \text{col}(X_a)$. Suppose we compare the models $EY \in X_a$ and $EY \in X_b$ using BIC. Get an exact expression for the probability that $X_b$ is favored over $X_a$ for a given sample size $n$, in the setting where the errors $\epsilon$ are Gaussian. Then give conditions for this probability to go to zero in a more general (non-Gaussian) setting.

3. Suppose we have a variable of interest $X_1$ and 10 possible confounders $X_2, \ldots, X_{11}$. All 11 variables are standardized. The possible confounders are exchangeably correlated with each other, meaning that $\text{cor}(X_j, X_k) = r_c$ for all $j \neq k \in 2, \ldots, 11$. The variable of interest has correlation $r$ with each of the possible confounders, so $\text{cor}(X_1, X_j) = r$ for $j = 2, \ldots, 11$. In the population mean structure

$$ E[Y|X] = \beta_0 + \sum_{j=1}^{11} \beta_j X_j, $$

half of the possible confounders have coefficient $\theta_c$ and half have coefficient zero. Suppose we use ridge regression with one of the two penalties:

- $\sum_{j=1}^{11} \beta_j^2$
- $\sum_{j=2}^{11} \beta_j^2$

Write code to numerically calculate the expected value of the GCV criterion as a function of $\lambda$ (note that this should not use simulation), and to identify the optimal value of $\lambda$. Then calculate the population MSE for $\hat{\beta}_1$ using this value of $\lambda$. Plot this MSE against $r$, $r_c$, $\theta_c$, and $\beta_1$. 

Stat 600 Problem Set 4
Due in class on 11/25.

1. (a) Suppose that $U$ and $V$ are bivariate Gaussian, and $g$ is a function. Show that

$$ \text{cov}(g(U), V) = \text{cov}(U, V) \text{cov}(g(U), U) / \text{var}(U). $$

(b) Suppose that we have a “single index model” in which

$$ Y = g(\alpha + \beta'X) + \epsilon, $$

Based on part (a), show that if $X$ is multivariate Gaussian with a non-singular covariance matrix, then OLS regression of $Y$ on $X$ yields $\hat{Y} = \hat{\alpha} + X \hat{\beta}$, where $\hat{\beta}$ consistently estimates $k\beta$ for a constant $k \in \mathbb{R}$.

(c) Use linearization to derive an approximate standard deviation for $\hat{\beta}_j/\hat{\beta}_k$ in the single index model.

(d) Use simulation to assess the estimates and standard errors when $g(z) = \text{arctan}(z)$.

2. Suppose we have two matrices $X_a$ and $X_b$, where $X_a$ is $n \times p + 1$, $X_b$ is $n \times p + q$, and $\text{col}(X_a) \subset \text{col}(X_b)$. We observe $Y = EY + \epsilon$, where $EY \in \text{col}(X_a)$. Suppose we compare the models $EY \in X_a$ and $EY \in X_b$ using BIC. Get an exact expression for the probability that $X_b$ is favored over $X_a$ for a given sample size $n$, in the setting where the errors $\epsilon$ are Gaussian. Then give conditions for this probability to go to zero in a more general (non-Gaussian) setting.

3. Suppose we have a variable of interest $X_1$ and 10 possible confounders $X_2, \ldots, X_{11}$. All 11 variables are standardized. The possible confounders are exchangeably correlated with each other, meaning that $\text{cor}(X_j, X_k) = r_c$ for all $j \neq k \in 2, \ldots, 11$. The variable of interest has correlation $r$ with each of the possible confounders, so $\text{cor}(X_1, X_j) = r$ for $j = 2, \ldots, 11$. In the population mean structure

$$ E[Y|X] = \beta_0 + \sum_{j=1}^{11} \beta_j X_j, $$

half of the possible confounders have coefficient $\theta_c$ and half have coefficient zero. Suppose we use ridge regression with one of the two penalties:

- $\sum_{j=1}^{11} \beta_j^2$
- $\sum_{j=2}^{11} \beta_j^2$

Write code to numerically calculate the expected value of the GCV criterion as a function of $\lambda$ (note that this should not use simulation), and to identify the optimal value of $\lambda$. Then calculate the population MSE for $\hat{\beta}_1$ using this value of $\lambda$. Plot this MSE against $r$, $r_c$, $\theta_c$, and $\beta_1$. 

1