Likelihood Theory and Extensions (BIOS:7110)
Breheny

Assignment 4
Due: Monday, September 19

1. *Gaussian graphical model.* As we discussed in class, the precision matrix $\mathbf{\Theta}$ is of interest as it describes conditional independence relationships. One way to estimate $\mathbf{\Theta}$ is $\widehat{\mathbf{\Theta}} = \mathbf{S}^{-1}$, where $\mathbf{S}$ is the sample variance-covariance matrix (throughout this problem, you may assume that all variance-covariance matrices are full rank). Here, we consider a different approach.

   (a) Suppose we partition $\mathbf{\Theta}$ so that the top left corner is isolated (i.e., the top left corner of the partition is 1×1 and the bottom right is $(d-1) \times (d-1)$, where d is the dimension of the multivariate distribution). Show that

   $$-\boldsymbol{\theta}_{21}/\theta_{11} = \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21},$$

   where $\mathbf{\Sigma}$ is the variance-covariance matrix, partitioned in the same way as $\mathbf{\Theta}$. Hint: use the definition of a matrix inverse.

   (b) Now consider the conditional distribution of $x_1|\mathbf{x}_2$. Show that if $\mathbf{x}$ is multivariate normal, then the conditional distribution of $x_1|\mathbf{x}_2$ can be written as

   $$X_1 = \alpha + \mathbf{x}_2^\top \boldsymbol{\beta} + \varepsilon,$$

   where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Express $\boldsymbol{\beta}$ and σ^2 in terms of the precision matrix $\mathbf{\Theta}$.

   (c) Part (b) suggests that we can estimate $\mathbf{\Theta}$ using linear regression. Simulate some multivariate normal data using the following code:

   ```
   set.seed(1)
   n <- 100
   A <- rnorm(n)
   B <- A + rnorm(n)
   C <- B + rnorm(n)
   D <- B + rnorm(n)
   X <- cbind(A, B, C, D)
   S <- cov(X)
   ```

   Then regress each element of $\mathbf{x}$ on the others. We are going to use these regression fits to estimate $\mathbf{\Theta}$; however, let us carry out a simple model selection procedure first, in which we drop any covariates that are not significant at the $\alpha = 0.05$ level. Then refit the model with only the significant covariates, and use $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ to fill in the appropriate elements of $\mathbf{\Theta}$; set $\beta_j = 0$ if the term was not included in the model. As an answer, you only need to provide $\hat{\mathbf{\Theta}}$, not the full summary of all the regression fits.

   (d) Does your estimate of $\mathbf{\Theta}$ from (c) reflect the correct conditional independence relationships among A, B, C, and D? Comment briefly.

   (e) Letting $\mathbf{x}^\top = [A\,B\,C\,D]$, show that the data generating mechanism of the above code results in $\mathbf{x}$ having a multivariate normal distribution, and calculate the true precision matrix $\mathbf{\Theta}^*$.

(f) We now have two estimators for $\boldsymbol{\Theta}$: $\mathbf{S}^{-1}$ and the estimator from part (c). Which one is more accurate (for this particular data set)? Quantify the overall accuracy using $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F$.

(g) One downside of the approach in (c) is that the estimate it produces, $\widehat{\boldsymbol{\Theta}}$, is asymmetric. One simple remedy is to use $\widetilde{\boldsymbol{\Theta}} = \frac{1}{2}\widehat{\boldsymbol{\Theta}} + \frac{1}{2}\widehat{\boldsymbol{\Theta}}^\top$ instead. Does this symmetrized estimate improve accuracy?

2. *Power calculation using the noncentral χ^2 distribution.* Suppose there is a latent random variable of interest Z that is continuously distributed between 0 and 1, but we observe only which of 10 bins it falls into: $(0, 0.1), (0.1, 0.2), \ldots, (0.9, 1.0)$. Thus, we observe $\mathbf{x}$, a 10-dimensional random vector of counts corresponding to the bins, with n denoting the total count. This problem involves attempting to test the null hypothesis that all bins are equally likely by assuming that $\mathbf{x}$ (approximately) follows a multivariate normal distribution.

   (a) Using the mean and variance of a multinomial distribution under the null, provide a function of $\mathbf{x}$ that follows an approximate χ^2 distribution (i.e., that would follow a χ^2 distribution if $\mathbf{x}$ were multivariate normal with the specified mean and variance).

   (b) Now suppose that $Z \sim \text{Beta}(1, 2)$. Create a plot overlaying two beta distributions: this one and the one corresponding to the null hypothesis.

   (c) Under the alternative distribution specified in (b), the quantity from (a) will no longer follow an ordinary χ^2 distribution, but instead a noncentral χ^2 distribution. Create a plot overlaying two χ^2 densities, one of the null hypothesis and the other with a noncentrality parameter of 10. Use the number of degrees of freedom appropriate to this problem.

   (d) Derive the noncentrality parameter for the distribution of $\mathbf{x}$ under the alternative hypothesis. Note that there is a problem with this calculation, in that the alternative hypothesis affects both the mean and the variance. For the purposes of this calculation, only account for its effect on the mean – assume that the variance is unchanged. Implement this calculation in a function, `ncp(n)`; turn this code in separately as a `.R` file so that I can run it and see that it works correctly.

   (e) Create a plot of n versus power (assuming an $\alpha = 0.05$ significance threshold), where n ranges from 10 to 100. Note that this calculation uses both the null distribution from (a) and the alternative distribution you derived in (d).

   (f) In the power calculation above, there are two potential issues: (a) the true distribution is multinomial, not multivariate normal, and (b) we ignored the impact of the alternative distribution on variance when calculating the noncentrality parameter. Carry out a simulation to compare the true power to our approximation. Draw samples of size $n = 50$ from the multinomial distribution and carry out the χ^2 test that you derived above (don't "correct" for continuity). Calculate the average power over $N = 10,000$ replications.

   (g) Comment briefly on the how the two approaches compare. In particular, suppose you were had to perform a power calculation like this for a real-world project: which approach would you use? Why?

3. *Bounded in probability vs convergence in distribution.* Prove that if a sequence of random vectors $\mathbf{x}_n \in \mathbb{R}^d$ converges in distribution, then $\mathbf{x}_n$ is bounded in probability.