HGEN 48600/STAT 35450: Fundamentals of Computational Biology

Lecture 14 — February 18

Lecturer: Matthew Stephens *Scribe: Lauren E. Blake*

Note: These lecture notes are still rough, and have only have been mildly proofread.

14.1 Class notes

1. There are two vignettes on the multivariate normal on Matthew's website:
   https://github.com/stephens999/fiveMinuteStats/tree/gh-pages/analysis
2. The reading for this lecture can be found in Chapter 10 of the Ross textbook.

14.2 Background

If X_1 and X_2 are independent normals (~ N(0, 1)), then aX_1 + bX_2 ~ N(0, $a^2 + b^2$)

14.3 Introduction to the multivariate normal

- There are many ways to definite the multivariate normal, but the definition from Wikipedia states that Vector X = (X_1, ... X_r) is r-variate normal or "r-dimensional multivariate normal" if every alternate linear combination of X, $\lambda_1 X_1 + \ldots + \lambda_r X_r$, is univariate normal (as long as $\lambda \neq 0$).
- Suppose Z = (Z_1, ... Z_n) are iid ~ N(0,1). Let A by an r x n matrix and μ be an r-vector. Then X = μ + AZ is a multivariate normal with mean E(X_j) = μ_j and covariance matrix Cov(X_i, X_j) = (A A′)$_{ij}$.
- For example, Cov(X_1, X_1) = Var(X_1) so the diagonal of the covariance matrix is one.

14.4 An example

There is an example in the vignette, Z_1, Z_2, Z_3 where

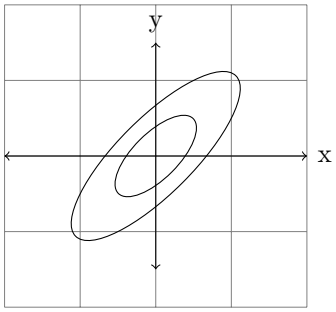
$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and where $X_1 = Z_1$ and Z_2 and $X_2 = Z_1 + Z_3$.

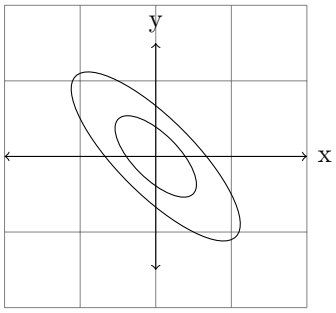
From this, we get X = AZ $\sim$ N(0, Σ = A A$'$)

- Note that Σ = A A$'$ =
$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$
- The covariance comes from the fact that they share 2 elements (Z_1) and don't share two elements (Z_2 and Z_3).
- Corr =
$$\frac{Cov}{\sqrt{(Var)}\sqrt{(Var)}}$$
=
$$\frac{1}{2}$$
=
$$\frac{Cov}{\sqrt{(2)}\sqrt{(2)}}$$
.

  This gives us a 3D ellipse that is tilted towards the right:



- If we were to change the covariance matrix so A A$'$ =
$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$
then we would have an 3D ellipse with the same shape but tilted to the left:



14.5 Simulating from the multivariate

- We can use the fact that a multivariate normal are linear combinations of univariate normals
- For $N_r(\mu, \Sigma_r)$, you can simulate from this provided that you can find A such that A $A' = \Sigma$.
- Note: Σ must be symmetric to hold. This means that Σ must be positive and semi-definite (so it has all non-negative eigenvectors).
- Inverting matrices can be unstable, so we need at least one other way to find the inverse of a matrix. In the Cholesky decomposition of Σ in the lower triangular matrix L has a value such that L $L' = \Sigma$. This allows use to find L and the from that find Z and then find X. We can then add μ if we are so inclined.
- To find the inverse of a matrix in R, use the command *chol2 inv (chol(Σ))*.
- It is important to note that we use this because inverting matrices can be unstable.
- We don't always have a density but if it is invertable (has all positive eigenvalues), then the density of the multivariate normal is the following:

$$p(x) = (\frac{1}{\sqrt{(2\pi)|\Sigma|}})^{\frac{r}{2}} exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)) \tag{14.1}$$

14.5.1 Maximizing p(x) for the multivariate normal

- It is not straightforward to show how to maximize p(x).
- If $X_1 \ldots X_n$ (that are iid) $\sim N_r(\mu, \Sigma)$, then the MLE for μ and Σ is $\hat{\mu}_{ij} = \frac{1}{n} \sum_i x_{ij}$. This is the average of vectors.
- Here, we are going to average j matrices: $\hat{\Sigma} = \frac{1}{n} \sum_i (\underline{x_i} - \underline{\hat{\mu}})(\underline{x_i} - \underline{\hat{\mu}})'$
- Therefore, $\hat{\Sigma} = \frac{1}{n} \sum_i (x_{ij} - \hat{\mu}_j)(\underline{x_{ik}} - \hat{\mu}_k)$
- The likelihood:

$$L(\mu, \Sigma) = \prod_i (\frac{1}{\sqrt{(2\pi)|\Sigma|}})^{\frac{r}{2}} exp(-\frac{1}{2}(x_i - \mu)'\Sigma^{-1}(x_i - \mu)) \tag{14.2}$$

14.6 Introduction to Gaussian Processes and Brownian Models

- If X_1, X_2 ... is a Markov chain, then $X_{t+1} \mid X_t \sim$ N(X_t, 1). It is a normal Markov chain where be begin at the origin (0,0) and $X_1 \sim$ N(0,1). As a result, we might call the a random walk, where each time we go an amount from a normal distribution.
- If we were to stop at X = 1000, then it would be a 1000-D multivariate normal where

  $X_1 = Z_1$
  $X_2 = X_2 + Z_1 = Z_1 + Z_2$
  $X_3 = X_2 + Z_3 = Z_1 + Z_2 + Z_3$

  so $\underline{X} = AZ$.
- Note: With a multivariate normal, they are uncorrelated if and only if they are independent.
- In the example $\underline{X} = AZ$, the points closer together are going to appear more correlated than points that are farther away.

- $\Sigma = \text{A A}' =$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

- But this is really interesting what you take the inverse $\Omega = \Sigma^{-1} =$

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

  where Ω is the precision matrix.

- The structure of the inverse of the covariance matrix can indicate that this is a Markov chain because they have the following special property:

  1. If X is multivariate normal, $\underline{\text{X}} \sim N_r(0, \Omega^{-1})$
  2. $\Omega^{ij} = 0$ if and only if X_i and X_j are conditionally independent given all other Xs
  3. $\Sigma_{ij} = 0$ if and only if X_i and X_j are independent

14.7 (Undirected) Gaussian graphical models

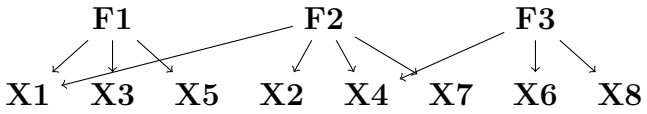
14.7.1 Introduction

- For any precision matrix, you want to make a connection if there is a non-zero covariance. This has many applications, including a gene network. For any precision matrix, you would make a connection between genes that have a non-zero covariance.

- The advantage of Gaussian graphical models is that you decrease the number of parameters that you are trying to estimate

- In the gene network example, you may have hub genes or master regulators that control a bunch of other genes. The signal from the controlled genes are observed, but the signal from the hub genes are not observed. Given that these are unobserved, there are no conditional independents. For example, the precision matrix for X_1, X_2, ... (observed signals) is dense but the precision matrix with F_1 ... (from the unobserved signals)

and X_1 ... (from the observed signals) is sparse. Removing an unobserved value won't drastically change a covariance matrix (it will change at only 1 row), but it will change the precision matrix everywhere.



- It is appealing for genetics and other applications, however, we are plagued by noisy data (e.g. expression data from RNA-seq). Since the true value of the measurement is unknown, we have to treat everything as unobserved, which is problematic.

14.7.2 Utilization

- Assume the observations have come to a relatively small number of factors so that the relationships between the observations are relatively simple.

$$\mathbf{X} = LF + E \tag{14.3}$$

where L represents the loadings, F represents the factors, and E represents the error terms. **X** has dimensions p x n. L has dimensions p x k and F has dimensions k x n where k is a small number of factors. E has dimensions p x n.

- In a genetics application, F is individual transcription factors, where each TF has different effects across the observations (e.g. TF binding sites near genes).

- If we assume that k factors F are independent, $N_n(0, \text{I})$ and rows of E are independent $N_n(0, \psi)$, then $X_i \sim \text{N}(0, \text{L L}' + \psi)$. In words, X_i is represented by a multivariate normal with rank k matrix and diagonal, ψ.

- We are making 3 assumptions in order to decrease the number of parameters to be estimated:

  1. The covariance matrix is sparse.
  2. The inverse covariance matrix is sparse.
  3. The covariance matrix is rank k normal with a diagonal. This is also known as factor models