HGEN 48600/STAT 35450: Fundamentals of Computational Biology

Lecture 15 — February 23

Lecturer: Matthew Stephens *Scribe: Nan Xiao*

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

This lecture discussed more properties and applications of multivariate normal distributions, and introduced Gaussian processes (Brownian motion) with examples.

15.1 Multivariate normal distributions

15.1.1 Sparse factor models

From last lecture, we know the MLE of multivariate normals (a high-dimensional covariance matrix estimation problem) gives us too many parameters to estimate, so we need some assumptions to reduce the number of parameters. Here we briefly review the three methods for such purposes:

1. Assumption of sparse covariance matrix
2. Assumption of sparse precision matrix (mostly appears in undirected Gaussian graphical models, which leads to the important concept of conditional independence)
3. Sparse factor models (low-rank factorization)

Sparse factor models:

$$\underbrace{Y}_{n\times p} = \underbrace{L}_{n\times k} \cdot \underbrace{F}_{k\times p} + \underbrace{E}_{n\times p} \tag{15.1}$$

This could be also named matrix factorization, since this model will factorize a matrix into two low-rank matrices. We assumed that

$$\begin{aligned} Y_{\cdot j} &\sim \mathcal{N}(0, LL' + \psi) && (15.2)\\ \text{if}\quad F_{\cdot j} &\sim \mathcal{N}(0, I) && (15.3)\\ E_{\cdot j} &\sim \mathcal{N}(0, \psi) && (15.4) \end{aligned}$$

Further more, there is an equivalence between two assumptions:

$$\Sigma \text{ (covariance matrix) is low rank and diagonal} \Leftrightarrow \Sigma^{-1} \text{ low rank and diagonal.} \tag{15.5}$$

15.1.2 Properties of multivariate normal distributions

Property 1. Linear Combination of MVNs

Sums and linear combinations of MVNs are still MVN.

If

$$\underbrace{X}_{r\times 1} \sim \mathcal{N}_r(\underbrace{\mu}_{r\times 1}, \underbrace{\Sigma}_{r\times r}) \tag{15.6}$$

then

$$\underbrace{A}_{p\times r}\underbrace{X}_{r\times 1} \sim \mathcal{N}_p(\underbrace{A\mu}_{p\times 1}, \underbrace{A\Sigma A^T}_{p\times p}) \tag{15.7}$$

Moreover,

$$\underbrace{A}_{p\times r}\underbrace{X}_{r\times 1} + \underbrace{b}_{p\times 1} \sim \mathcal{N}_p(\underbrace{A\mu + b}_{p\times 1}, \underbrace{A\Sigma A^T}_{p\times p}) \tag{15.8}$$

Imagine for a bivariate normal distribution, the term b will only shift the mean, it will not change the shape of the distribution (the covariance is kept the same).

Property 2. Conditional distributions

Given a MVN X, if we condition on a subset of X, say X', then $X|X'$ is still MVN, and only the mean is changed, the covariance will not change.

Also, if

$$\begin{pmatrix} \underbrace{X_1}_{r_1} \\ \underbrace{X_2}_{r_2} \end{pmatrix} \sim \mathcal{N}_{r_1+r_2}\left(\begin{pmatrix} \underbrace{\mu_1}_{r_1} \\ \underbrace{\mu_2}_{r_2} \end{pmatrix}, \begin{pmatrix} \underbrace{\Sigma_{11}}_{r_1} & \underbrace{\Sigma_{12}}_{r_2} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then $X_1|X_2 = a$ is MVN. In fact,

$$X_1|X_2 \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}). \tag{15.9}$$

where

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2) \tag{15.10}$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{15.11}$$

See the Wikipedia page on conditional distributions of MVNs (`https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions`) for details.

A regression analog of the above equation is, if:

$$Y = X\beta + E \tag{15.12}$$

and we know this is actually

$$E(Y|X) = X\beta \tag{15.13}$$

then

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2) \tag{15.14}$$

So $\Sigma_{12}\Sigma_{22}^{-1}$ is also called *regression coefficients*.

15.1.3 Application of MVN in HMM

Recall the Hidden Markov Model we used in the previous lectures (Figure 15.1):

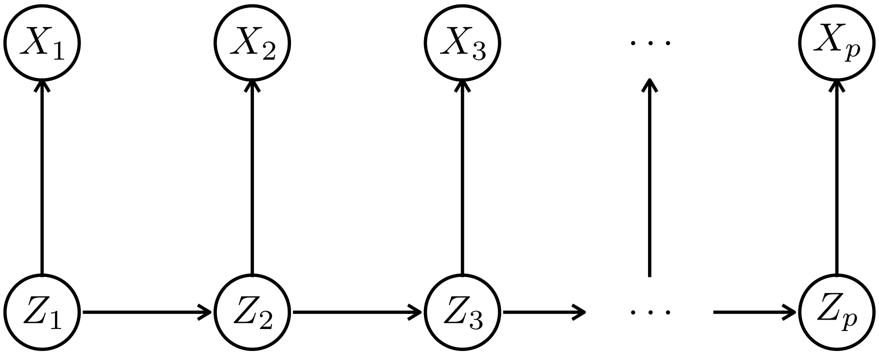

Figure 15.1. A Hidden Markov Model

Assume everything is MVN, suppose

$$\begin{aligned} Z_{t+1}|Z_t &\sim \mathcal{N}_r(Z_t, \Sigma_1) && (15.15)\\ X_t|Z_t &\sim \mathcal{N}_r(Z_t, \Sigma_2) && (15.16) \end{aligned}$$

If Z_i and X_i are r-dimensional vectors, then $X_1, X_2, \ldots, X_p$ and $Z_1, Z_2, \ldots, Z_p$ are two $p \times r$ multivariate normals.

Equation 15.14 is useful in forwardbackward algorithms, since the algorithm involves multivariate normal calculations, and particularly *conditional distributions* of $X_1|X_2 = a$ is MVN. Additionally, Kalman filter is also related to this property.

15.1.4 Linear algebra tricks to compute $\bar{\mu}$

Trick 1. Avoid computing matrix inverse directly

$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$. This formula involves computing the matrix inverse Σ_{22}^{-1} numerically. In matrix computations, we must avoid computing the inverse directly, instead, we should use matrix decomposition to do it.

For example, if $y = Ax$, we want to get x by solving $x = A^{-1}y$, but it turns out we don't need to solve A^{-1} directly. In reality, it is possible to use multiple types of matrix decompositions to get x, here we use Cholesky decomposition to do this.
Cholesky decomposition factorizes a matrix into the product of a lower triangular matrix and its conjugate transpose, namely:

$$A = LL' \tag{15.17}$$

then we have

$$y = LL'x \Rightarrow L'x = L^{-1}y \tag{15.18}$$

Since L^{-1} is lower triangular, we can get L^{-1} here easily by using the backsolve algorithm (the corresponding R function is `backsolve()`), then we get $L^{-1}y$, use backsolve again to get x easily, since L' is upper triangular.

Cholesky decomposition exists when A is a positive semi-definite (PSD) matrix, so it is very useful in factorizing covariance matrices, since covariance matrices are always PSD.

Trick 2. Avoid re-computing matrix inverses

If we already knew what A^{-1} is (e.g. already computed it somehow), then $(A+uv')^{-1}$ is easy to compute, by using the *Sherman–Morrison formula*, and not compute it directly. Here uv' represents rank-1 changes to A, but actually this works on any low-rank modifications of A. We can use *Sherman–Morrison formula* to compute the rank-1 updates of A:

$$(\underbrace{A}_{n\times n} + \underbrace{u}_{n\times 1}\underbrace{v'}_{1\times n})^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u} \tag{15.19}$$

The *Sherman-Morrison-Woodbury formula* is an natural extension for the above formula, on rank-k modification of A, i.e.

$$(A + UV)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1} \tag{15.20}$$

One example of this is factor models. In fact, $LL' = l_1 l_1' + l_2 l_2' + \cdots$, equivalent to changing k columns in L.

Another example is variable selection regressions. If we have the linear model

$$\underbrace{Y}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{E}_{n\times p} \tag{15.21}$$

we can use MCMC to fit it, and we assume that β is sparse (only a few of the coefficients are non-zero). This will require changing one β_i from 0 to non-zero, which is a small modification to the original matrix.

Trick 3. Avoid computing ratios directly by division

In Metropolis–Hastings algorithm, we have to compute the ratio of two densities:

$$\frac{\pi(x')}{\pi(x)} \tag{15.22}$$

In this case, we should compute the logarithm of the two densities and then exponentiate the difference between them, instead of computing the division directly, i.e.

$$\exp(\log\pi(x') - \log\pi(x)) \tag{15.23}$$

since $\pi(x')$ and $\pi(x')$ are close to 0, the division between two near 0 numbers will be numerically unstable.

For multivariate normals, we should also do this:

$$\cdots \propto \frac{1}{|\Sigma^{p/2}|}\exp(\cdots) \tag{15.24}$$

When we compute likelihood ratios, we should also compute the two log-likelihoods, then do the exponentiation, rather than dividing two likelihood directly.

15.2 Brownian motion

Read Ross book Chapter 10 (Brownian Motion and Stationary Processes) for more details about this section.

Considering a symmetric random walk, which in each time unit we take a step either to the left or to the right. Suppose for each Δt time unit we take a step of size Δx either to the left or the right with equal probabilities. If $X(t)$ is the position at time t, then

$$X(t) = \Delta x(X_1 + \cdots + X_{[t/\Delta t]}) \tag{15.25}$$

where

$$X_i = \begin{cases} +1, & \text{if } i\text{-th step of length } \Delta x \text{ is to the right} \\ -1, & \text{if } i\text{-th step of length } \Delta x \text{ is to the left} \end{cases} \tag{15.26}$$

and $[t/\Delta t]$ represents the largest integer less than or equal to $t/\Delta t$, and X_i are independent,

$$P(X_i = 1) = P(X_i = -1) = \frac{1}{2}. \tag{15.27}$$

This is essentially a Markov chain with $P_{i,i+1} = P_{i,i-1} = 1/2, i = 0, \pm 1, \ldots$

Suppose that we take smaller and smaller steps in smaller and smaller time intervals. If we go to the limit in the right manner, what we obtain is defined as *Brownian motion.*

Take the limit, we get:

$$X(t) \sim \mathcal{N}(0, \sigma^2 t) \tag{15.28}$$

When $\sigma = 1$, the process is called *standard Brownian motion* (we can get a standard Brownian motion by standardizing any Brownian motion by $B(t) = X(t)/\sigma$), then

$$X(t) \sim \mathcal{N}(0, t) \tag{15.29}$$

Independent Increment Property:

$$X(t) - X(s) \sim \mathcal{N}(0, (t-s)), \quad \forall s < t. \tag{15.30}$$

Recall the normal Markov Chain simulations we did in the past lectures (`http://stephens999.github.io/fiveMinuteStats/analysis/normal_markov_chain.html`):

$$\begin{aligned} X_1 &\sim \mathcal{N}(0,1) \\ X_2 &= X_1 + \mathcal{N}(0,1) \\ X_3 &= X_2 + \mathcal{N}(0,1) \\ &\ldots \end{aligned}$$

Intuitively, this property implies such a continuous Markov chain with $\mathcal{N}(0,(t-s))$ embedded in it.

15.2.1 Application of Brownian motion in genetics and evolution

Figure 15.2 shows an example of Brownian motion on a species tree of certain trait (e.g. height). Assume we start from $t = 0$, and $z_i \sim \mathcal{N}(0, t_i)$. We can see that the bottom nodes of the tree are linear combinations of z_i.

Since $z_i \sim \mathcal{N}(0, t_i)$, we know that $\sum z_i$ are multivariate normals, with covariance matrix associated with the length of the changes.

Such evolution of traits can be also modelled using the *Ornstein–Uhlenbeck processes*, which introduces certain strength of attraction for each step, in addition to the simple Brownian motion model.

For more details of the model, search “Brownian motion model of trait evolution”.

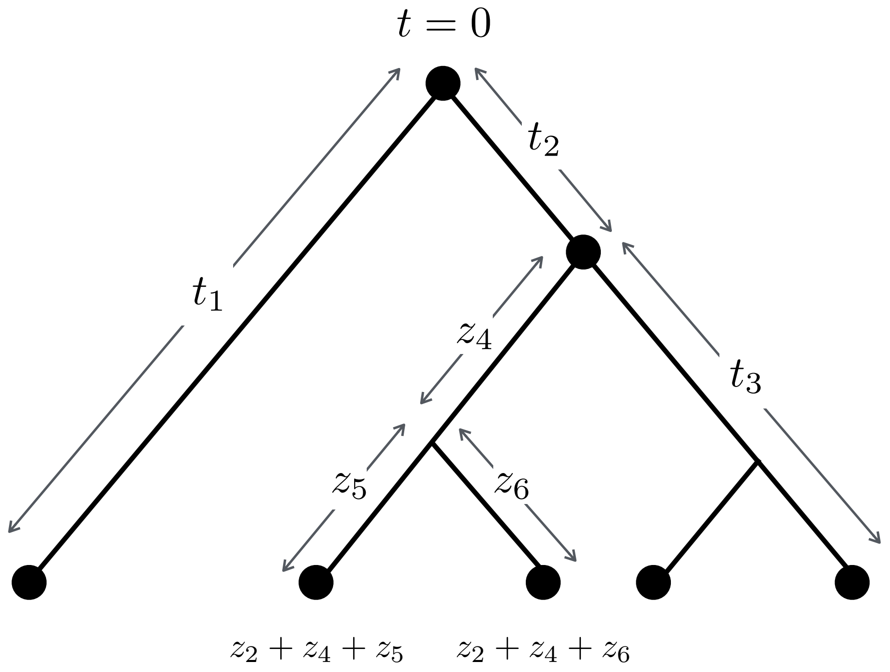

Figure 15.2. Brownian motion along an evolution tree

15.2.2 General definition of a Gaussian Process

Definition. A time-continuous stochastic process is a Gaussian Process $\iff$ For every set of indices in the index set T, $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ is multivariate Gaussian.

As a matter of fact, continuous time can always be discretized. Usually, the covariance matrix is proportional to the time length (or we can somehow know their definition of pairwise covariances); the mean is always 0. Therefore, a Gaussian Process is often defined by specifying $\text{Cov}(X(t_1), X(t_2)), \forall\, t_1, t_2$.

We usually use K to denote this function:

$$K(t_1, t_2) = \text{Cov}(X(t_1), X(t_2)). \tag{15.31}$$

If $K(t_1, t_2)$ only depends on $|t_1 - t_2|$, then it is said to be *stationary*.

This indicates that the nearer the two points are, the more correlated they will be. The farther the two points are, the less correlated they will be, as shown in Figure 15.3.

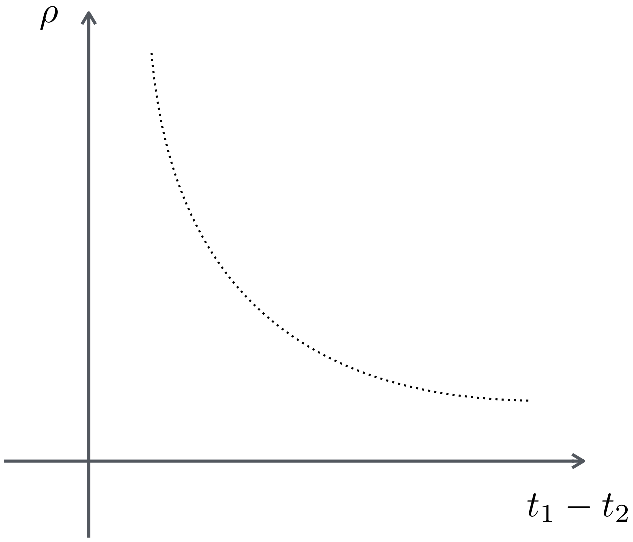

Figure 15.3. The relationship between time length and correlation of two points.

This implies the *smoothness* for the "local" time points, and the *scattered* pattern for the points in the "global" perspective.

In fact, most processes we use are stationary. Brownian motion is one exception.