Review of distribution theory

Anirban Bhattacharya

http://www.stat.tamu.edu/~anirbanb

January 18, 2018

1 Positive definite matrices: review of some basic facts

For $x \in \mathbb{R}^d$, let $\|x\| = \sqrt{\sum_{j=1}^d x_j^2}$ denote its Euclidean norm. Also, for $x, y \in \mathbb{R}^d$, let $\langle x, y\rangle := x^{\mathrm{T}} y$ denote the inner product between x and y. Recall the Cauchy–Schwartz inequality $\langle x, y\rangle \leq \|x\| \, \|y\|$.

A $d \times d$ matrix Ω is called positive definite (p.d.) if Ω is symmetric and $x^{\mathrm{T}}\Omega x > 0$ for all $x \neq 0 \in \mathbb{R}^d$. Ω is called positive semi-definite (p.s.d.) or non-negative definite (n.n.d.) if Ω is symmetric and $x^{\mathrm{T}}\Omega x \geq 0$ for all $x \neq 0 \in \mathbb{R}^d$.

Note: some definitions of p.d. and p.s.d. matrices do not require symmetry as I do above.

Eigen-decomposition: For any p.s.d. matrix Ω, there exist orthogonal vectors $v_1, \ldots, v_d \in \mathbb{R}^d$ satisfying $\|v_j\| = 1$ and $\langle v_j, v_{j'}\rangle = 0$ for all $j \neq j'$, and non-negative numbers $\lambda_1 \geq \ldots \lambda_d \geq 0$, such that

$$\Omega = \sum_{j=1}^d \lambda_j v_j v_j^{\mathrm{T}}.$$

The above decomposition is the eigen-decomposition of Ω; the vectors v_js are the eigenvectors and λ_js the corresponding eigenvalues, with $\Omega v_j = \lambda_j v_j$ for all j. Any Ω as above is clearly p.s.d., as $x^{\mathrm{T}}\Omega x = \sum_{j=1}^d \lambda_j \langle v_j, x\rangle^2$. Hence, we have a characterization of p.s.d. matrices.

Letting V denote the $d \times d$ matrix whose columns are v_j, and Λ the diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_d$, we can write the identity in the display more concisely as $\Omega = VDV^{\mathrm{T}}$. The matrix V is an orthogonal matrix, satisfying $VV^{\mathrm{T}} = V^{\mathrm{T}}V = \mathrm{I}_d$. Using the identities $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, and $\det(AB) = \det(BA)$, it is also clear from the above decomposition that $\mathrm{tr}(\Omega) = \sum_{j=1}^d \lambda_j$, and $\det(\Omega) = \prod_{j=1}^d \lambda_j$.

Cholesky decomposition: If Ω is p.d., then there exists a unique lower triangular matrix L with positive diagonal entries such that $\Omega = LL^{\mathrm{T}}$. The Cholesky decomposition is typically less expensive to compute than the eigendecomposition. We shall see applications later on.

2 Review of some continuous distributions

Any course on Bayesian statistics cannot even begin without a good dose of distribution theory since all uncertainty statements in Bayesian statistics is quantified through probability distributions. I am going to lay down conventions/notations used throughout the course, and also some warmup exercises. Some of these will be posted in homework0, however, you are encouraged to work through the details whenever I leave something for you to verify (VFY!). Familiarity with these basic distributional manipulations will be assumed and heavily used.

Exponential distribution. Say $X \sim \mathrm{Expo}(\lambda)$ (Exponential distribution with rate parameter

λ) if X has pdf

$$f(x) = \lambda e^{-\lambda x}, \quad x \in (0, \infty).$$

The mean of the distribution is $1/\lambda$ and variance $1/\lambda^2$.

Gamma distribution. Say $X \sim \text{Gamma}(\alpha, \beta)$ (Gamma distribution with shape parameter α and rate parameter β) if X has pdf

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}, \quad x \in (0, \infty).$$

The mean is α/β and variance α/β^2. Caveat: Lots of places (books, softwares, etc.) use the shape-scale formulation, with the scale being the inverse-rate β^{-1}. Stick to one convention to avoid confusion.

Chi-square distribution. The χ^2 distribution with ν degrees of freedom ($\nu > 0$) is the Gamma density with shape $\nu/2$ and rate $1/2$.

Inverse-Gamma distribution. Say $X \sim \text{Inv-Gamma}(\alpha, \beta)$ (Inverse Gamma distribution with shape parameter α and rate parameter β) if X has pdf

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta/x} x^{-(\alpha+1)}, \quad x \in (0, \infty).$$

The mean is $\beta/(\alpha - 1)$ for $\alpha > 1$.

Beta distribution. $X \sim \text{Beta}(a, b)$ has pdf

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in (0, 1).$$

Gaussian/Normal distribution. We shall use $N(\mu, \sigma^2)$ to denote the normal distribution with mean μ and variance σ^2. We shall often encounter normal distributions in the following form in this course: we obtain a distribution which we know only up to some constant of proportionality,

$$f(x) \propto e^{-\frac{1}{2}(qx^2 - 2bx)}, \quad x \in \mathbb{R},$$

where $q > 0$ and $b \in \mathbb{R}$ are parameters of the distribution. The quadratic form (with leading constant positive) inside the negative exponent tells us this is a normal distribution. Moreover, some algebra tells us the mean is $q^{-1}b$ and variance is q^{-1} (VFY!) What is the normalizing constant (i.e., constant of proportionality)?

We shall often use ϕ and Φ to denote the pdf and cdf of the standard normal distribution, i.e., $\mu = 0$ and $\sigma = 1$.

Truncated Gaussian distribution. The truncated Gaussian distribution with parameters μ, σ, restricted to the interval (u, v), has density,

$$f(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbb{1}_{(u,v)}(x),$$

where $\mathbb{1}_A$ denotes the indicator function of set A. Can you find the normalizing constant of this distribution, in terms of Φ? (Start with the standard case).

In general, if f is a pdf on $\mathbb{R}$, we shall call the pdf

$$g(x) = \frac{f(x)\, \mathbb{1}_{(u,v)}(x)}{\int_u^v f(z)dz}$$

to be f truncated to (u, v).

Student's t distribution A random variable $X \sim t_\nu$ (t distribution with $\nu > 0$ degrees of freedom) has density

$$f(x) = \frac{\Gamma(\nu + 1/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\,\frac{1}{(1 + x^2/\nu)^{(\nu+1)/2}}, \quad x \in \mathbb{R}.$$

When $\nu = 1$, this is the standard Cauchy density. The mean exists for $\nu > 1$ and equals zero. The variance exists for $\nu > 2$ and is $\nu/(\nu - 2)$.

For $\mu \in \mathbb{R}$ and $\sigma > 0$, we say that $Z \sim t_\nu(\mu, \sigma^2)$ if $(Z - \mu)/\sigma \sim t_\nu$. We shall call μ the center and σ the scale. The general t family is a location-scale family.

The following result is an important one. The t distribution can be expressed as a variance mixture of a Gaussian. Specifically, suppose $X \mid \tau \sim N(\mu, \tau^{-1}\,\sigma^2)$, and $\tau \sim \text{Gamma}(\nu/2, \nu/2)$. Then, $X \sim t_\nu(\mu, \sigma^2)$ (VFY!).

This is an example of a hierarchical formulation. This is a much more general idea, where a distribution is specified hierarchically, by first defining a distribution given some parameters, and then assigning those parameters additional distributions. For example, we first define $X \mid \tau \sim f(\cdot \mid \tau)$, where f is a density which depends on parameter τ in some way, and then assign $\tau \sim p$. (Note: Although in the previous example τ was a scalar, in general, it could be a vector as well) Then, the marginal density of X is

$$f_X(x) = \int f(x \mid \tau)\, p(\tau) d\tau.$$

We shall see that hierarchical modeling plays an extremely important role in Bayesian statistics.

We have seen that for $\nu > 2$, the variance of $t_\nu(\mu, \sigma^2)$ is $\{\nu/(\nu - 2)\}\sigma^2 > \sigma^2$. Without actually calculating the variance, can you argue that the variance of t will be larger (if it exists)?

2.1 Multivariate distributions

We shall mainly encounter two multivariate distributions in this course, the multivariate Gaussian/normal (MVN) and the Dirichlet distribution.

MVN distribution. The d-variate normal distribution $\mathcal{N}_d(\mu, \Sigma)$ has pdf

$$f(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\, e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d,$$

with mean $\mu \in \mathbb{R}^d$, and $d \times d$ positive definite (p.d.) covariance matrix Σ. The following results are useful:

Theorem.

(i) If X has a d-variate normal distribution, then any $k < d$ marginal of X is also normal.

(ii) Let $X \sim N_d(\mu, \Sigma)$. Let A be $k \times d$ ($k \leq d$) full rank and b be $k \times 1$. Then $AX + b \sim N_k(A\mu + b, A\Sigma A^{\mathrm{T}})$.

(iii) Let us partition

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $X^{(1)}$ is $d_1 \times 1$ and $X^{(2)}$ is $d_2 \times 1$ with $d_1 + d_2 = d$. Similarly for μ and Σ. Note $\Sigma_{21} = \Sigma_{12}^{\mathrm{T}}$. The conditional distribution of $X^{(1)} \mid X^{(2)}$ is

$$N\big(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(X^{(2)} - \mu^{(2)}), \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big).$$

As in the univariate case, we shall often encounter the MVN distribution in the form,

$$f(x) \propto e^{-\frac{1}{2}(x'Qx - 2b'x)}, \quad x \in \mathbb{R}^d,$$

where Q is a p.d. matrix. The quadratic form in the negative exponent tells us the joint distribution is Gaussian, with mean $Q^{-1}b$ and covariance matrix Q^{-1} (VFY!). Q is called the inverse covariance or precision matrix, and is a key object for multivariate modeling, especially graphical models.

Dirichlet distribution $X = (X_1, \ldots, X_d) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d, \alpha_{d+1})$ has pdf

$$f(x_1, \ldots, x_d) = \frac{\Gamma(\sum_{j=1}^{d+1} \alpha_j)}{\prod_{j=1}^{d+1} \Gamma(\alpha_j)} x_1^{\alpha_1 - 1} \ldots x_d^{\alpha_d - 1}(1 - x_1 - \ldots - x_d)^{\alpha_{d+1} - 1}, \quad x_i \in (0,1) \,\forall\, i, \; \sum_{i=1}^{d} x_i \leq 1.$$

The Dirichlet distribution is a multivariate generalization of the Beta distribution. The following characterization is super useful: if $T_i \sim \Gamma(\alpha_i, 1)$ independently for $i = 1, \ldots, d+1$, and $T = \sum_{i=1}^{d+1} T_i$, then

$$(T_1/T, \ldots, T_d/T) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d, \alpha_{d+1}).$$

Thus, if we normalize independent Gammas with the same rate parameter, we get a Dirichlet distribution. If you have not seen this before, good to do this calculation using the multivariate change of variable theorem. This property is useful in proving various stuff about the Dirichlet distribution, for example, that lower-dimensional marginals of a Dirichlet are also Dirichlet (with what parameters?).