Notes for Building Statistical Software

Robert deCarvalho

Contents

Chapter 1

Introduction

This document is a collection of various statistical equations that should serve as a useful resource for building statistical software. In particular, it should be a useful reference for maintaining the django-brain software package created for my current employer, Ambition.

It is assumed throughout, that the reader has some familiarity with the statistical concepts and notation involved. Think of this document as a collection of statistical facts that have been assembled from various sources, or derived where necessary.

Chapter 2

Normal Distributions

2.1 Density Functions

The univariate probability density function is

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

The univariate cumulative density function is

$$F(x|\mu,\sigma) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{2.2}$$

The multivariate probability density function is given below. In these equations, bold-face type indicates either a column vector (lower-case) or a matrix (upper-case).

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\left(\frac{k}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\ &= \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{(2\pi)^{\left(\frac{k}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})}\end{aligned} \tag{2.3}$$

Where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, and k is the dimensionality of the vector space.

2.2 Adding Distributions

The sum of N normally distributed random variables is also normally distributed with parameters

$$\boldsymbol{\mu} = \sum_{i=1}^{N} \boldsymbol{\mu}_i \qquad \text{and} \qquad \boldsymbol{\Sigma} = \sum_{i=1}^{N} \boldsymbol{\Sigma}_i \tag{2.4}$$

2.3 Multiplying Distributions

The product of N normal distributions is also a normal distribution with parameters

$$\boldsymbol{\Lambda} = \sum_{i=1}^{N} \boldsymbol{\Lambda}_i \qquad \text{and} \qquad \boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1} \sum_{i=1}^{N} \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i \tag{2.5}$$

Where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix.

2.4 Incrementally Incorporating Data Into Distributions

This section builds on a white paper by Raul Rojas entitled "The Kalman Filter." We'll start by considering the univariate normal distribution and then generalize to the multivariate case.

Assume you start with some initial values for μ_0 and σ_0. The following recursive update equations can be used to incrementally update these initial values to reflect all the values observed thus far.

$$\begin{aligned} \mu_{n+1} &= \frac{n}{n+1}\mu_n + \frac{1}{n+1}x_{n+1} \\ \sigma^2_{n+1} &= \frac{n}{n+1}\sigma^2_n + \frac{n}{(n+1)^2}\left(x_{n+1} - \mu_n\right)^2 \end{aligned} \tag{2.6}$$

The following equivalent formulation is sometimes easier to think about and will lead us towards handling the multivariate case.

$$\begin{aligned} K &\equiv \frac{1}{n+1} \\ \mu_{n+1} &= \mu_n + K\left(x_{n+1} - \mu_n\right) \\ {\sigma'}^2_n &= \sigma^2_n + K\left(x_{n+1} - \mu_n\right)^2 \\ \sigma^2_n &= (1-K){\sigma'}^2_n \end{aligned} \tag{2.7}$$

The equations in (2.7) can be generalized to obtain the updated multivariate mean, $\boldsymbol{\mu}_{n+1}$ and covariance, $\boldsymbol{\Sigma}_{n+1}$.

$$\begin{aligned} K &\equiv \frac{1}{n+1} \\ \boldsymbol{\mu}_{n+1} &= \boldsymbol{\mu}_n + K\left(\boldsymbol{x}_{n+1} - \boldsymbol{\mu}_n\right) \\ \boldsymbol{\Sigma}'_n &= \boldsymbol{\Sigma}_n + K\left(\boldsymbol{x_{n+1}} - \boldsymbol{\mu_n}\right)\left(\boldsymbol{x_{n+1}} - \boldsymbol{\mu_n}\right)^T \\ \boldsymbol{\Sigma}_{n+1} &= (1-K)\boldsymbol{\Sigma}'_n \end{aligned} \tag{2.8}$$

2.5 Incremental Updates with "Forgetting"

In this section, I describe a method for incrementally incorporating data into a normal distribution while discounting older data. This is very similar to what a Kalman filter does, only quite a bit more simple.

The update equations (2.7) (2.8) in the previous section contain the "gain" factor, K, which depends on how many data points the distribution has already ingested. The gain factor is essentially a weighting factor indicating the relative importance of the current data point with respect to all previously observed points. Now consider what happens if we initialize our distribution with some μ_0 and σ_0 and artificially freeze the number of points to some value of our choosing, n_{max}. We then continually feed the distribution identical data points with some final value x_f and watch how the mean value changes. The gain factor, of course, will be constant at

$$K_{max} = \frac{1}{1 + n_{max}}. \tag{2.9}$$

Then starting with the second line of (2.7) we see

$$\begin{aligned} \mu_{n+1} &= \mu_n + K_{max}\left(x_{n+1} - \mu_n\right) \\ &= \mu_n(1 - K_{max}) + K_{max}x_f \end{aligned} \tag{2.10}$$

and defining $A \equiv 1 - K_{max}$ and $B \equiv K_{max}x_f$ we see that the update equation takes on a form which should be familiar to anyone who has examined the amortization problem for compounded interest.

$$\mu_{n+1} = A\mu_n + B. \tag{2.11}$$

To see the correspondence, imagine that n indexes the number of months you are into your mortgage, μ_n represents the outstanding balance on your mortgage, A represents your monthly interest rate, and B represents your monthly payment (with a negative sign).

2.5.1 Exploring the Amortization Update Equation

The amortization equation can apply to many situations. Although we are specifically interested in applying it to the "forgetting" problem, we could apply it to a variety of problems. We therefore reexpress the equation in terms of some general variable x with the understanding that we can apply it to our specific case of μ whenever we want.

$$x_{n+1} = Ax_n + B. \tag{2.12}$$

Let's assume we start off with some initial value for x that we will call x_0, and then step through the first few iterations of equation (2.12).

$$
\begin{aligned}
x_0 &= x_0 \\
x_1 &= Ax_0 + B \\
x_2 &= A^2x_0 + AB + B \\
x_3 &= A^3x_0 + AB^2 + AB + B \\
&... \\
x_n &= A^nx_0 + B\sum_{i=0}^{n-1} A^i
\end{aligned} \tag{2.13}
$$

The summation in (2.13) occurs quite frequently in different analysis problems, and it has a closed-form solution as long as $A < 1$:

$$\sum_{i=0}^{n-1} A^i = \frac{1 - A^N}{1 - A}. \tag{2.14}$$

We use this result to obtain the final result for our amortization problem.

$$
\begin{aligned}
x_n &= A^nx_0 + \frac{B}{1-A} - \frac{BA^n}{1-A} \\
&= A^n\left(x_0 - \frac{B}{1-A}\right) + \frac{B}{1-A}
\end{aligned} \tag{2.15}
$$

Notice that (2.15) is exponential in n. This means that the value will start at x_0 and exponentially approach the final value $B/(1 - A)$. It is often more convenient to think about exponential processes in terms of their exponential "decay constant" which we will label as n_d and compute with the following steps.

$$
\begin{aligned}
A^n &= e^{-n/n_d} \\
ln(A) &= -\frac{1}{n_d} \\
n_d &= -\frac{1}{ln(A)}
\end{aligned} \tag{2.16}
$$

This means that equation (2.15) can also be expressed as

$$x_n = e^{-n/n_d}\left(x_0 - \frac{B}{1-A}\right) + \frac{B}{1-A} \tag{2.17}$$

2.5.2 Applying Amortization Results to the "Forgetting Problem"

So let's pick up where we left of with the problem of incrementally updating a normal distribution while "forgetting" old observations. We had a recursive update equation for the mean value given by

$$\mu_{n+1} = A\mu_n + B.$$

where

$$\begin{aligned} A &= 1 - K_{max} \\ B &= x_f k_{max} \\ k_{max} &= \frac{1}{1 + n_{max}} \end{aligned} \tag{2.18}$$

I won't go through all the steps here, but applying the amortization results to these equations will give you the following result.

$$\mu_n = e^{-ln\left(\frac{n_{max}+1}{n_{max}}\right)n}(x_0 - x_f) + x_f \tag{2.19}$$

So let's say in words what we have shown here. Placing an upper limit on the number of points a normal distribution thinks it has observed will have the following effects. As long as the number of observations is less than the limit, the normal distribution will behave perfectly "normal" and continually become more and more precise in its estimate of the mean. As soon as the threshold is crossed though, it will subtly change its behavior and begin to follow the general trends of the input data maintaining an estimate of the current uncertainty. It will exponentially forget older measurements with a decay constant given by equation (2.16). Although derived for the univariate case, similar arguments can be made for the multivariate case.

One other point of interest involves the decay constant which is given by

$$n_d = \frac{1}{\ln\left(\frac{n_{max}+1}{n_{max}}\right)}. \tag{2.20}$$

This can be inverted to get

$$n_{max} = \frac{1}{e^{1/n_d} - 1}. \tag{2.21}$$

This means that if you want your response to have a decay constant of n_d observations, you should set n_{max} according to (2.21).

2.6 Conditioning

One of the most common problems in statistics is to determine causation. Basically people want to be able to answer the question "If I change this variable, how likely is it that this

other variable will change, and by how much?" In general, causation is hard to prove, but one ingredient for causation is correlation. If movement in one variable induces movement in another, then they will be correlated. The multivariate distribution provides some nice tools for analyzing these correlations.

Recall that the multivariate normal distribution has a density function given by

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\left(\frac{k}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{2.22}$$

Evaluation this function at a particular value for $\boldsymbol{x}$ will tell you how likely you are to observe a set of variables with values near $\boldsymbol{x}$. But you can then ask the question, "If I know the value of the first element of $\boldsymbol{x}$, how does that change the distribution for all the rest of the values in $\boldsymbol{x}$? This is called conditioning.

We will study conditioning by partitioning the elements of $\boldsymbol{x}$ into two sets, the unknown set $\boldsymbol{x_1}$ and the known set $\boldsymbol{x_2}$. What this means is that we know the values for $\boldsymbol{x_2}$ and want the updated probability distribution for the unknown set $\boldsymbol{x_1}$. Let's see how this works. First I carry out my partitioning

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \tag{2.23}$$

This then leads to corresponding partitions for the mean vector

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \tag{2.24}$$

and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{2.25}$$

Ignoring the normalization constant, this means I can rewrite the probability density function as

$$\mathcal{P}(\boldsymbol{x_1},\boldsymbol{x_2}) \sim e^{-\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}}. \tag{2.26}$$

But what I really want to know is the conditional distribution

$$\mathcal{P}(\boldsymbol{x}_1|\boldsymbol{x}_2 = \boldsymbol{a}). \tag{2.27}$$

The algebra required to get this result is super tedious and involves something called the Schur complement, so I'm just going to state the result. The conditional distribution is just another multivariate normal,

$$\mathcal{P}(\boldsymbol{x}_1|\boldsymbol{x}_2 = \boldsymbol{a}) = \mathcal{N}\left(\boldsymbol{x}_1|\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\Sigma}}\right) = \frac{1}{|\bar{\boldsymbol{\Sigma}}|^{\frac{1}{2}}(2\pi)^{\left(\frac{k}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x_1}-\bar{\boldsymbol{\mu}})^T\bar{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x_1}-\bar{\boldsymbol{\mu}})} \tag{2.28}$$

where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{a} - \boldsymbol{\mu}_2\right) \tag{2.29}$$

and

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{2.30}$$

2.7 Marginalizing

Compared to conditioning, marginalizing is pretty simple. Let's go back to the partitioned multivariate normal we used above.

$$\mathcal{P}\left(\boldsymbol{x_1}, \boldsymbol{x_2}\right) \sim e^{-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2\end{bmatrix}}. \tag{2.31}$$

Marginalizing means that you are answering the question "What is the probability density of $\boldsymbol{x}_1$ when the full distribution has been integrated over all possible values of $\boldsymbol{x}_2$?" The answer is super straightforward. The marginalized densities are just multivariate normals that depend only on their own partitions

$$\mathcal{P}\left(\boldsymbol{x}_1\right) = \mathcal{N}\left(\boldsymbol{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}\right) = \frac{1}{|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}(2\pi)^{\left(\frac{k_1}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)} \tag{2.32}$$

and

$$\mathcal{N}\left(\boldsymbol{x}_2\right) = \mathcal{N}\left(\boldsymbol{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}\right) = \frac{1}{|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}(2\pi)^{\left(\frac{k_2}{2}\right)}} e^{-\frac{1}{2}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)} \tag{2.33}$$

Chapter 3

Beta Distributions

3.1 Density Functions

The probability density function is

$$f\left(x|\alpha,\beta\right) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathcal{B}(\alpha,\beta)}$$

where the beta function, $\mathcal{B}(\alpha,\beta)$ serves as the normalizer.

The mean and variance are given by

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{3.1}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3.2}$$

These can be inverted to get the α and β parameters as a function of the mean and standard deviations. Note that this is not completely reversible in that it is possible to specify mean and standard deviation values that cannot be realized by the beta distribution.

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) \tag{3.3}$$

$$\beta = \left(\frac{1}{\mu} - 1\right) \tag{3.4}$$

Chapter 4

Gamma Distributions

4.1 Density Functions

The probability density function has two equivalent parameterizations

$$f\left(x|k,\theta\right) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \tag{4.1}$$

$$f\left(x|\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{4.2}$$

$$\tag{4.3}$$

The mean and standard deviations for the two parameterizations are

$$\mu = k\theta = \frac{\alpha}{\beta} \tag{4.4}$$

$$\sigma = \sqrt{k}\theta = \frac{\sqrt{\alpha}}{\beta} \tag{4.5}$$

For a given μ and σ, the parameters are

$$k = \left(\frac{\mu}{\sigma}\right)^2 \qquad \theta = \frac{\sigma^2}{\mu} \tag{4.6}$$

and

$$\alpha = \left(\frac{\mu}{\sigma}\right)^2 \qquad \beta = \frac{\mu}{\sigma^2} \tag{4.7}$$

Chapter 5

Log-Normal Distributions

5.1 Density Function

The log-normal density function is

$$f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{5.1}$$

The first and second moments are given by

$$m = \langle x \rangle = e^{\mu+\sigma^2/2} \tag{5.2}$$

$$s^2 = \langle x^2 \rangle = \left(e^{\sigma^2}-1\right) e^{2\mu+\sigma^2} = \left(e^{\sigma^2}-1\right) m^2 \tag{5.3}$$

The parameters expressed in terms of the moments are

$$\sigma^2 = \ln\left(1+\frac{\langle x^2\rangle}{\langle x\rangle^2}\right) = \ln\left(1+\frac{s^2}{m^2}\right)$$

$$\mu = \ln\langle x\rangle - \frac{1}{2}\sigma^2 = \ln m - \frac{1}{2}\sigma^2 \tag{5.4}$$

5.2 Sum of Log-Normal Variables

Let's say I want to sum N variables, each of which is taken from the same log-normal distribution with first and second moments m and s^2 respectively. I want know the distribution of that sum. There is no exact closed-form solution to this problem, but I can construct a reasonable approximation as follows. The first moment of the sum will just be the sum of the first moments, and the second moment of the sum will just be the sum of the second moments, or

$$M = Nm \qquad\qquad S^2 = Ns^2. \tag{5.5}$$

I can then compute the parameters corresponding to these moments and assume that the final distribution will also have the form of a log-normal. Note that this is not rigorously correct and is an approximation. Nevertheless, continuing on, I can substitute equations (5.5) into (5.4) to arrive at the log-normal parameters for the sum distribution

$$\begin{aligned} \sigma_s^2 &= \ln\left[1 + \frac{1}{N}\left(e^{\sigma^2} - 1\right)\right] \\ \mu_s &= \ln\left(Ne^{\mu}\right) + \frac{1}{2}\left(\sigma^2 - \sigma_s^2\right). \end{aligned} \tag{5.6}$$