Multivariate Methods

Homework #1 (2018-2019), Answers

Q1: Another Lagrange Multiplier application

(As you may well know) the entropy of a discrete distribution is given by $H(\vec{p}) = -\frac{1}{\log 2}\sum_i p_i \log p_i$. *Consider a discrete distribution in which* p_n *is the probability of drawing a token of "value"* n *, where* $n = 0,1,2,3,\cdots$. *Find the distribution* $\vec{p}$ *that maximizes* $H(\vec{p})$ *subject to the constraint that the average value is equal to* A *, i.e., that* $\sum_{i=0}^{\infty} ip_i = A$.

It suffices to maximize $(\log 2)H(\vec{p}) = -\sum_i p_i \log p_i$. There are two constraints: that the probability distribution sums to 1, i.e., that $\sum_{i=0}^{\infty} p_i = 1$, and that $\sum_{i=0}^{\infty} ip_i = A$. We associate each constraint with its own Lagrange multiplier, and find extrema of the unconstrained problem by setting derivatives to zero.

$L(\vec{p}) = (\log 2)H(\vec{p}) + \lambda\sum_{i=0}^{\infty} p_i + \mu\sum_{i=0}^{\infty} ip_i = -\sum_i p_i \log p_i + \lambda\sum_{i=0}^{\infty} p_i + \mu\sum_{i=0}^{\infty} ip_i$, so

$\frac{\partial}{\partial p_k} L(\vec{p}) = -\frac{p_k}{p_k} - \log p_k + \lambda + \mu k = -1 - \log p_k + \lambda + \mu k$.

We set the partial derivatives of $\frac{\partial}{\partial p_k} L(\vec{p}) = 0$:

$-1 - \log p_k + \lambda + \mu k = 0$, so $\log p_k = 1 - \lambda - \mu k$, i.e., $p_k = e^{1-\lambda-\mu k}$.

We now need to find the multipliers λ and μ so that the constraints are satisfied.

For normalization (λ): $\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} e^{1-\lambda-\mu k} = e^{1-\lambda}\sum_{k=0}^{\infty} e^{-\mu k} = \frac{e^{1-\lambda}}{1-e^{-\mu}}$.

For the mean (μ):

$$\begin{aligned} \sum_{k=0}^{\infty} kp_k &= \sum_{k=0}^{\infty} ke^{1-\lambda-\mu k} = e^{1-\lambda}\sum_{k=0}^{\infty} ke^{-\mu k} = e^{1-\lambda}\left(\sum_{k=1}^{\infty} e^{-\mu k} + \sum_{k=2}^{\infty} e^{-\mu k} + \sum_{k=2}^{\infty} e^{-\mu k} + \cdots\right) \\ &= e^{1-\lambda}\left(\frac{e^{-\mu}}{1-e^{-\mu}} + \frac{e^{-2\mu}}{1-e^{-\mu}} + \frac{e^{-3\mu}}{1-e^{-\mu}}\cdots\right) = e^{1-\lambda}\frac{e^{-\mu}}{1-e^{-\mu}}\left(1 + e^{-\mu} + e^{-2\mu} + \cdots\right) = e^{1-\lambda}\frac{e^{-\mu}}{(1-e^{-\mu})^2} \end{aligned}$$.

So the constraints are satisfied if

$\frac{e^{1-\lambda}}{1-e^{-\mu}} = 1$ and $e^{1-\lambda}\frac{e^{-\mu}}{(1-e^{-\mu})^2} = A$. Substituting the first equation into the second yields $\frac{e^{-\mu}}{1-e^{-\mu}} = A$, so

$e^{-\mu} = A(1-e^{-\mu})$, and $e^{-\mu} = \frac{A}{1+A}$, and $1 - e^{-\mu} = 1 - \frac{A}{1+A} = \frac{1}{1+A}$

So $p_k = e^{1-\lambda-\mu k} = (1-e^{-\mu})e^{-\mu k} = \left(\frac{1}{1+A}\right)\left(\frac{A}{1+A}\right)^k = \frac{A^k}{(1+A)^{k+1}}$.

We also have to check that this extremum is maximum, not a minimum or saddle point. This follows from general properties of entropy, and the nature of the constraints. First we note that a mixture of two distributions that satisfy the constraints will also satisfy the constraints, since the constraints are linear in the distribution.

But also, since mixing distributions can only increase their entropy, any interior extremum must be a maximum. The maximum also must be a global maximum, since if there were two local maxima, then their mixture would have a still higher entropy. Finally, the extremum identified above is an interior point, because each of the p_k is nonzero.

Note that the above strategy is applicable for maximizing the entropy of a distribution subject to any set of lnear constraints – which include mean, variance, skewness, kurtosis, and correlation structure. However, the equations to satisfy the constraints may not be easily solvable.

Q2: Regression and "cross-correlation analysis" (from MVAR1415)
Consider the standard regression scenario described in the class notes, pages 1-2. That is, there are n observations, $y_1,\ldots,y_n$, and p regressors, where the typical regressor $\vec{x}_j$ is a column $x_{1,j},\ldots x_{n,j}$, and the set of p regressors forms a $n \times p$ matrix X, and we seek a set of p coefficients $b_1,\ldots,b_p$, the $p \times 1$ matrix B, for which $\left|Y - XB\right|^2$ is minimized.
Now let's assume that the regressors $\vec{x}_j$ are orthonormal. For example, we're doing spatial receptive field analysis. Here $x_{i,j}$ corresponds to the luminance presented on the ith trial in pixel j, and we've designed our stimuli so that, over the entire stimulus sequence, $\sum_{i=1}^{N} x_{i,j}x_{i,k} = 0$ if $j \neq k$, and $\sum_{i=1}^{N} x_{i,j}x_{i,j} = 1$.
How does this simplify the computation of the regressors B?

We have the formal solution $B = (X^*X)^{-1}X^*Y$. The assumptions of orthonormality, namely, $\sum_{i=1}^{N} x_{i,j}x_{i,k} = 0$ if $j \neq k$, and $\sum_{i=1}^{N} x_{i,j}x_{i,j} = 1$, mean that $(X^*X)_{j,k} = (X^TX)_{j,k} = \sum_i x_{i,j}x_{i,k}$. (Since the *x*'s are all real, $X^* = X^T$.) So X^*X is the identity matrix, and $B = (X^*X)^{-1}X^*Y = X^*Y$. That is, the model coefficients *B* can be computed by correlating the response sequence *Y* against the stimulus sequences *X*; no matrix inversion is needed.

An extension of this argument – choosing the regressors to be a sequence for that is orthogonal to time-shifts of itself (e.g., an m-sequence) leads to the "reverse correlation" procedure for determining the temporal aspects of receptive fields.