

**Alternative conceptual models of maximum entropy estimators
of probability mass functions: primal, dual and triple.**

D R A F T

Please do not quote. Comments and corrections welcome.

November 8, 2004

by

Robert A. Collins
Naumes Professor
OMIS Department
Santa Clara University

Alternative conceptual models of maximum entropy estimators of probability mass functions: primal, dual and triple.

0. Introduction

Estimators for maximum entropy probability mass functions are an application of standard constrained optimization tools using the method of Lagrange. As with any constrained optimization problem, it may be viewed as finding a saddle point and Farkas's lemma guarantees it is the maximum of the primal model and the minimum of the dual model. Where \mathbf{P} is an n dimensional vector of probabilities and $f(\mathbf{P})$ is the objective function to be maximized subject to the m constraints $g_k(\mathbf{P}) = 0$, where $m < n$, the general form of the primal model is:

$$\max_{w.r.t. P} L(P, \lambda) = f(P) + \sum_k \lambda_k g_k(P)$$

The general form of the dual model is:

$$\min_{w.r.t. \lambda} L(P, \lambda) = f(P) + \sum_k \lambda_k g_k(P)$$

$$\text{subject to: } \frac{\partial f}{\partial p_i} + \sum_{k=1}^m \lambda_k \frac{\partial g_k}{\partial p_i} = 0 \quad k = 1, \dots, m$$

In the maximum entropy setting, $f(\mathbf{P})$ is the Shannon measure of uncertainty and the $g_k(\mathbf{P})$ are moment conditions. The primal problem is interpreted as finding the probabilities that put the maximum amount of uncertainty in the mass function, and still have it meet the moment restrictions. Therefore, the probability mass function reflects only the information contained in the moment requirements, and no more. The optimized value of the Lagrangian L^* is the maximum uncertainty that can exit in the mass function if it reflects only the information that is contained in the moment requirements. This information theory interpretation of these models is widely understood. The lambdas reflect the marginal increase in uncertainty that must be reflected in the mass function as the moments increase causing more dispersion to be required in the data. While some regard the primal and the dual as two different conceptualizations of the problem, they may also be thought of as two different ways to represent the same problem.

This paper, while building on the achievements of Zellner and Highfield, Golan Judge and Miller and others, presents an entirely new way of conceptualizing the estimation of maxent probability mass functions. Instead of using constrained optimization, the method outlined here collapses the problem into a single equation and the solution is found by finding a tangent hyperplane to a strictly convex function. The strictly convex function may be conceptualized as the moment generating function of the mass function being estimated, and the tangency may be thought of as finding the spot on the moment generating function where the moment conditions are satisfied. This conceptualization produces the identical set of gradient conditions as the standard saddle point problem. Therefore for this particular optimization problem, in addition to the primal and the dual conceptualization of the problem, there is also a "triple" model that has an entirely different conceptualization, but produces exactly the same solution.

1. The maximum entropy primal model

The objective is to maximize Shannon's uncertainty measure,

$$-\sum_{i=1}^n (p_i) \ln(p_i)$$

subject to $m+1$ arithmetic moment constraints [around the origin] of the form:

$$\sum_{i=1}^n x_i^k p_i = \mu_k \text{ for } k = 0, \dots, m \quad \mu_0 = 1$$

where the p_i are the probabilities to be calculated, the x_i are the pre-specified support points for the probability mass function, and the μ_k are the required values for the moments. The dimension of the problem is determined by k , the number of moment restrictions, and the number of points chosen for the mass function (n) may be as large as desired and not create a computational burden. The Lagrangian objective function for the maxent problem with $n+m+1$ variables has the form:

$$L(p_1, \dots, p_n, \lambda_0, \dots, \lambda_m) = -\sum_{i=1}^n (p_i) \ln(p_i) + \lambda_0 \left(\sum_{i=1}^n x_i^0 p_i - 1 \right) + \dots + \lambda_m \left(\sum_{i=1}^n x_i^m p_i - \mu_m \right)$$

The first-order condition for a maximum requires $n+m+1$ partial derivatives. The first n f.o.c. equations for the probabilities are:

$$\frac{\partial L}{\partial p_i} = -\ln(p_i) - \frac{1}{p_i}(p_i) + \lambda_0 + x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m = 0.$$

If the requirement of the zero moment [summing constraint] is ignored, these equations can be solved producing n equations of the form:

$$p_i = e^{\lambda_0 - 1 + x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} = e^{\lambda_0 - 1} \left[e^{x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} \right].$$

Differentiating L with respect to the lambdas produces the next $m+1$ equations of the first-order conditions which are the moment conditions.

$$\begin{aligned} \frac{\partial L}{\partial \lambda_0} &= \sum_{i=1}^n x_i^0 p_i - 1 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= \sum_{i=1}^n x_i^1 p_i - \mu_1 = 0 \\ \frac{\partial L}{\partial \lambda_2} &= \sum_{i=1}^n x_i^2 p_i - \mu_2 = 0 \\ &\dots \\ \frac{\partial L}{\partial \lambda_m} &= \sum_{i=1}^n x_i^m p_i - \mu_m = 0 \end{aligned}$$

The typical method of solution is to substitute the solutions for the first n equations for the p_i into the constraint equations reducing the system of equations to $m+1$ non-linear equations which, following Zellner and Highfield, may be solved with a Taylor series expansion or by application of a standard gradient method.

Following Golan, Judge and Miller, the system of $m+1$ equations may be reduced to m equations by scaling the individual probabilities to sum to one thus eliminating the need for the summing constraint. Dividing each probability by the sum of the n probabilities scales them to sum to one. The sum of the n probabilities is:

$$\sum_{i=1}^n P_i = e^{\lambda_0 - 1 + x_1 \lambda_1 + x_1^2 \lambda_2 + \dots + x_1^m \lambda_m} + e^{\lambda_0 - 1 + x_2 \lambda_1 + x_2^2 \lambda_2 + \dots + x_2^m \lambda_m} + \dots + e^{\lambda_0 - 1 + x_n \lambda_1 + x_n^2 \lambda_2 + \dots + x_n^m \lambda_m}$$

$$= e^{\lambda_0 - 1} \left[e^{x_1 \lambda_1 + x_1^2 \lambda_2 + \dots + x_1^m \lambda_m} + e^{x_2 \lambda_1 + x_2^2 \lambda_2 + \dots + x_2^m \lambda_m} + \dots + e^{x_n \lambda_1 + x_n^2 \lambda_2 + \dots + x_n^m \lambda_m} \right]$$

Dividing each probability by the sum, the equations for the scaled probabilities are:

$$P_i(\lambda_1, \dots, \lambda_m) = \frac{e^{x_i \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m}}{e^{x_1 \lambda_1 + x_1^2 \lambda_2 + \dots + x_1^m \lambda_m} + e^{x_2 \lambda_1 + x_2^2 \lambda_2 + \dots + x_2^m \lambda_m} + \dots + e^{x_n \lambda_1 + x_n^2 \lambda_2 + \dots + x_n^m \lambda_m}}$$

which is frequently written:

$$P_i = \frac{1}{\Omega} \left[e^{x_i \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} \right]$$

where omega is:

$$\Omega(\lambda_1, \lambda_2, \dots, \lambda_m) = e^{x_1 \lambda_1 + x_1^2 \lambda_2 + \dots + x_1^m \lambda_m} + e^{x_2 \lambda_1 + x_2^2 \lambda_2 + \dots + x_2^m \lambda_m} + \dots + e^{x_n \lambda_1 + x_n^2 \lambda_2 + \dots + x_n^m \lambda_m}$$

Omega is often referred to as a “partition function” but as we will see, it also has another interpretation.

2. The maximum entropy dual model

The general form of the dual model is:

$$\min \text{ w.r.t } \lambda \quad L(P, \lambda) = f(P) + \sum_k \lambda_k g_k(P)$$

$$\text{subject to: } \frac{\partial f}{\partial p_i} + \sum_k \lambda_k \frac{\partial g_k}{\partial p_i} = 0$$

The set of n dual constraints just require the n probabilities to take on their usual maximum entropy form. Differentiating the Shannon measure $[f(\mathbf{P})]$ with respect to p_i produces $-\ln(p_i) - 1$ and the derivative of the k^{th} moment constraints $[g_k(\mathbf{P})]$ with respect to p_i is x_i^k .

Therefore the n dual constraints are of the form:

$$-\ln(p_i) - 1 + \lambda_0 + x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m = 0 \text{ for } i = 1, \dots, n$$

Solving for p_i , we see the n probabilities take on their usual exponential form:

$$p_i = e^{\lambda_0 - 1 + x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} = e^{\lambda_0 - 1} \left[e^{x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} \right] \text{ for } i = 1, \dots, n$$

Again, the probabilities can be scaled to leave out the summing constraint. Whether scaled or not, they can be simply substituted back into the objective function which makes everything a function of the lambdas. By substituting the constraints into the objective function, the problem becomes unconstrained and can be minimized w.r.t the m or m+1 lambdas with m or m+1 terms in the sum.

$$\min \text{ w.r.t } \lambda: L[P(\lambda), \lambda] = f[P(\lambda)] + \sum_i \lambda_i g_i[P(\lambda)]$$

The minimum value of L will be where the sum has a value of zero. Since the lambdas are not zero, the sum is minimized where the g_i functions are all zero. This produces the same set of m [or m+1] equations that must be solved in the primal:

$$\begin{aligned} \sum_{i=1}^n x_i^1 p_i - \mu_1 &= 0 \\ \sum_{i=1}^n x_i^2 p_i - \mu_2 &= 0 \\ &\dots \\ \sum_{i=1}^n x_i^m p_i - \mu_m &= 0 \end{aligned}$$

3. The maximum entropy triple model

While all constrained optimization problems have a primal and a dual, for this problem there is a third completely different ‘triple’ model that produces the identical first-order conditions. We want to estimate a maximum entropy probability mass function subject to moment restrictions. Starting with the moment generating function of the mass function to be created:

$$M(t) = E\{e^{tX}\} = \sum_{i=1}^n e^{tx} f(x_i) = \sum_{i=1}^n e^{tx} p_i$$

In the context of max entropy, however, the form of f(x) is: [using the scaled version]

$$p_i = \frac{1}{\Omega} \left[e^{x_i^1 \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} \right]$$

Therefore, the moment generating function for a mass function meeting the criterion of maximum entropy with moments of any permissible value may be written:

$$M(t) = \frac{1}{\Omega} \sum_{i=1}^n e^{tx_i} \left\{ e^{x_i \lambda_1 + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m} \right\}$$

$$= \frac{1}{\Omega} \sum_{i=1}^n e^{x_i(t + \lambda_1) + x_i^2 \lambda_2 + \dots + x_i^m \lambda_m}$$

As usual for any moment generating function, the first partial of M with respect to t where t = 0 produces the first moment, etc. But in this case, the moments can be defined by differentiating the moment generating function w.r.t **either** t or the lambdas where the kth partial with respect to t equals the derivative with respect to the kth lambda when t is set equal to zero.



Therefore, in lambda space, the moments of the mass function being estimated are:

$$\sum_{i=1}^n x_i^k p_i = \frac{1}{\Omega} \frac{d\Omega}{d\lambda_k}$$

This means that in the context of maximum entropy estimation, the natural log of partition function may be regarded as the moment generating function in lambda space.

$$\frac{\partial^k M(t)}{\partial t^k} = M_k = \frac{d \ln \Omega(\lambda)}{d \lambda_k}$$

At all points on the moment generating function, the maximum entropy criterion is satisfied, but the moments vary according to the [m dimensional] slope at that point. This means that in this setting, the maximum entropy problem may be regarded as searching the slope of the moment generating function to find the points where the moments have the required values. Therefore, the goal is not to maximize or minimize, but to find a tangency to the moment generating function. This m dimensional tangency is where the gradient of the moment generating function equals the vector of the required values for the moments. The gradient vector of the moment generating function is:

$$\nabla M(t) = \left[\frac{\partial M(t)}{\partial \lambda_1}, \frac{\partial M(t)}{\partial \lambda_2}, \dots, \frac{\partial M(t)}{\partial \lambda_m} \right]$$

By specifying a 1 by m vector of the required values of the moments = μ

$$\mu = [\mu_1, \mu_2, \dots, \mu_m]$$

the solution to the maximum entropy estimation reduces to a single equation:

$$\nabla M(t)(\lambda_1, \dots, \lambda_m) = \mu.$$

Therefore, the maximum entropy estimation may also be conceptualized as finding a m-dimensional tangency to the moment generating function. Since the derivatives of the moment generating function returns the moments, this equation reflects the same system of equations as the primal model and the dual model.

$$\sum_{i=1}^n x_i^1 P_i - \mu_1 = 0$$

$$\sum_{i=1}^n x_i^2 P_i - \mu_2 = 0$$

...

$$\sum_{i=1}^n x_i^m P_i - \mu_m = 0$$

4. Concluding remarks

The standard calculations for estimating a probability mass function that meets the maximum entropy criterion arises from three conceptual models. In addition to the usual primal and dual models, the identical calculations arise from conceptualizing the process as finding a point of tangency to a moment generating function.

References

Golan, A., G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, Chichester, UK: John Wiley and Sons, 1996.

Zellner, A. and R. Highfield, "Calculation of Maximum Entropy Distributions and Approximation of Marginal Posterior Distributions," *Journal of Econometrics*, 1988 37, 195-209.