PRELIMINARY, INCOMPLETE AND FOR DISCUSSION ONLY

Extracting infromation from observed counts

Avinash Singh Bhati*

October 9, 2004

Abstract

This paper describes an information-theoretic approach for analyzing count outcomes. It discusses extensions to the standard Poisson regression model that incorporate such real-world features as overdispersion, abnormally inflated counts, truncation & censoring, and nesting with minimal reliance on distributional assumptions. At the conference, several examples, using real-world data sets, will be used to demonstrate the information-recovering potential of the framework.

1. INTRODUCTION

In applied work, researchers are often confronted with outcomes that appear as non-negative integers. Such outcomes could be modeled in the standard linear framework either in their manifest forms or after some transformation (e.g., logarithmic, freeman-tukey, etc.). Recognizing the discrete nature of the outcomes, their highly skewed nature, the presence of one or more abnormally frequent outcomes, and other such peculiarities, such an approach

Contact Address: abhati@ui.urban.org, 202-261-5329, Justice Policy Cente, Urban Institute, 2100 M Street, NW, Washington, DC 20037. Part of this research was supported by grant 2002-IJ-CX-0006 from the National Institute of Justice. Opinions expressed in this document are those of the authors, and do not represent the official position or policies of the U.S. Department of Justice, the Urban Institute, its trustees, or its funders. I thank Dr. Amos Golan for helful comments and encouragement. All errors are mine alone.

is usually considered sub-optimal and researchers typically rely on Poisson based regressions or some variants thereof.

Regression analysis of count data are an extensively researched and written about topic in econometrics and several excellent reviews and texts exist (Cameroon and Trivedi, 1998; Winkelmann 2003). However, there are several features of count data models that are currently subjects of ongoing research. These included the analysis of truncated counts and censored samples, modeling the effects of endogenous regressors, the issue of unobserved heterogeneity (nesting effects) as well as extensions of the modeling framework to small samples. In addition, there are constant innovations to the modeling of extra-Poisson variation (overdispersion) in count outcome models and models that account for the preponderence of some outcomes (e.g., zero inflated count outcome models or hurdle models).

In this paper, we approach all these models in an information-theoretic framework and demonstrate the flexibility of the approach when an analyst is faced with one or more of these real-world challenges.

This paper is organized as follows: In the next section (TO COME), we give an overview of the count outcome setting and provide a concrete statement of the problem an econometrician faces. In the section following that, we provide a detailed description of how the information-theoretic approach can be applied to the count data setting. We begin by make some simple assumptions that mimick the assumptions underlying the Poisson distribution. Not surprisingly, it is seen there that these assumptions yield information-theoretic solutions identical to those obtained under the Maximum Likelihood estimation and inference framework applied to the Poisson model. That section, however, provides a slightly different way of motivating and deriving Poisson models and therefore suggests various places where assumptions could be relaxed and flexibility introduced. In the section following that we explicitly deal with one extension at a time by relaxing some assumtions

and/or adding futher structure. Each of these modifications yields a model that has one or more of these real-world features that one may wish to permit.

At the conference, we will demonstrate the information-recovery potential of the framework by applying it to several real-world data sets. We will also provide some (limited) comparisons with other familiar model. This paper does not provide any simulation results. That work is ongoing.

2. BACKGROUND

(TO COME)

3. SETTING UP THE BASIC PROBLEM

Consider, as a point of departure, that a set of N signals $(s_1, \ldots, s_N)$ are emitted from some source (nature, society, experimental apparatus, etc.) that we do not observe directly. Instead, we only have available imperfect manifestations of these signals in the form of N outcomes $(y_1, \ldots, y_N)$. Assume also that theory provides us only weak and partial guidance about the possible predictors of the signals and we wish to utilize all this knowledge to recover information about the signals.

In order to proceed, let us first convert all unknowns into proper probabilities. To do so, let us define each signal as an expectation over a support space. That is, let

$$s_n = \mathbf{z}'\mathbf{p}_n = \sum_m z_m p_{mn} \qquad \forall n, \tag{3.1}$$

where $\mathbf{z}$ is an M dimensional signal support space and $\mathbf{p}_n$ are a set of M proper probabilities that, when applied to the support space, yield the expected outcome—the signal. For sig-

nals yielding count outcomes, the support space is readily defined as a sequence of possible counts $\mathbf{z} = (0, 1, 2, 3, \ldots, z_M)'$ that encompasses the realm of possibility (i.e., we assume z_M is large enough). No assumptions are made about the probabilities of interest other than that they are proper (i.e., $p_{mn} > 0 \; \forall n, m$ and $\sum_m p_{mn} = 1 \; \forall n$).

Next we incorporate knowledge about the exogenous predictors by requiring the signals to satisfy certain constraints. In order to do so, consider defining certain features of the signals. E.g., $\frac{1}{N} \sum_n x_{kn} s_n$ is the covariance between the kth predictor and the expected outcomes (the signals). A natural sample analogue to use for this feature would be the covariance between the predictors and the observed outcomes. That is, if the signals are to mimick the structure in the observed outcomes, then it is reasonable to assume that

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}' \mathbf{p}_n \qquad \forall k \tag{3.2}$$

This may exhaust all the knowledge we have about the process but it results in a system of $N \times M$ unknowns with only $K + N$ equations linking them—an ill-posed inversion problem. As such, there are an infinite number of solutions that could satisfy the constraints. Following the literature in information-theory, one way out of this seemingly ill-defined problem is to maximize the incertainty implied by the probabilities while requiring them to satisfy all known constraints (Jaynes 1957a; Jaynes 1957b).

Using Shannon's (1948) Entropy as the criterion to quantify uncertainty, the information recovery task gets mathematically formulated as the following constrained optimization problem:

$$\max_{\mathbf{p}} \quad H(\mathbf{p}) = -\sum_n \mathbf{p}_n' \log \mathbf{p}_n \tag{3.3}$$

subject to the K moment constraints (3.2) and the adding-up constraint ($\sum_m p_{mn} = 1 \; \forall n$).

If, in addition to the above pieces of information, we have some non-sample informa-

tion in the form of prior probabilities, defined over the same space as the posteriors, then we can formulate the information recovery task as a constrained *minimization* problem where, subject to sample moment and adding up constraints, we minimize the cross entropy or the Kullback-Leibler directed divergence between the posteriors and the priors. If we have prior probabilities p^0_{mn} then the problem is to

$$\min_{\mathbf{p}} \quad \mathscr{K}(\mathbf{p};\mathbf{p}^0) = \sum_n \mathbf{p}'_n \log\left(\mathbf{p}_n/\mathbf{p}^0_n\right) \tag{3.4}$$

subject to (3.2) and the adding up constraints. This problem can be solved analytically (up to a set of lagrange multipliers) by setting up the primal lagrangian function as

$$\begin{aligned}\mathscr{L} \quad = \quad & \sum_n \mathbf{p}'_n \log\left(\mathbf{p}_n/\mathbf{p}^0_n\right) + \sum_n \eta_n\left\{1 - \mathbf{1}'\mathbf{p}_n\right\} \\ & \sum_k \lambda_k\left\{\sum_n x_{kn}y_n - \sum_n x_{kn}\mathbf{z}'\mathbf{p}_n\right\}\end{aligned} \tag{3.5}$$

where $\{\lambda_k\}$ and $\{\eta_n\}$ are the sets of lagrange multipliers related to the moment and adding up constraints, respectively. Solving the first order conditions, we obtain solutions for the probabilities of interest in the form

$$\hat{p}_{mn} = \frac{p^0_m \exp(z_m \sum_k x_{kn}\hat{\lambda}_k)}{\sum_m p^0_m \exp(z_m \sum_k x_{kn}\hat{\lambda}_k)} = \frac{p^0_m \exp(z_m \mathbf{x}'_n\hat{\boldsymbol{\lambda}})}{\hat{\Omega}_n} \tag{3.6}$$

where non-zero priors and the exponential form of the probabilities ensure that they are non-negative and the partition function Ω_n ensures that the probabilities are proper (sum to one).

Different assumtion about the prior probabilities, however, lead to very different models. As noted by Masaumi (1993), the cross entropy solution results in the ML Poisson model *if* we assume that the prior probability for observing z_m counts is exactly $(z_m!)^{-1}$. To

see this, note that using prior probabilities of $p_m^0 = \frac{1}{z_m!}\ \forall m$ and assuming a large enough z_M, the general solution of (3.6) can be written as

$$\begin{aligned} p_{mn} &= \frac{\frac{1}{z_m!}\exp(z_m \mathbf{x}_n'\boldsymbol{\lambda})}{\sum_{m=0}^{\infty}\frac{1}{z_m!}\exp(z_m \mathbf{x}_n'\boldsymbol{\lambda})} = \frac{\frac{1}{z_m!}\exp(z_m \mathbf{x}'\boldsymbol{\lambda})}{\exp(\exp(\mathbf{x}_n'\boldsymbol{\lambda}))} \\ &= \frac{\exp(-\exp(\mathbf{x}_n'\boldsymbol{\lambda}))\exp(\mathbf{x}_n'\boldsymbol{\lambda})^{z_m}}{z_m!} \\ &= \frac{\exp(-\alpha_n)\alpha_n^{z_m}}{z_m!} \end{aligned} \tag{3.7}$$

which is the Poisson distribution with a log link function, i.e., $\alpha_n = \exp(\mathbf{x}_n'\boldsymbol{\lambda})$ or $\log \alpha_n = \mathbf{x}_n'\boldsymbol{\lambda}$.

Inserting the optimal solution (3.6) obtained above back into the primal constrained minimization problem (3.4), we can solve for the unconstrained dual version of the problem that is a function of the K lagrange multipliers λ_k. The dual objective function is

$$\mathscr{L}_* = \sum_{nk} \lambda_k x_{kn} y_n - \sum_n \log \Omega_n \tag{3.8}$$

where Ω_n is the partition function defined above. The dual objective function typically does not have an analytical solution but a numeric one can be obtained in a variety of software that permit non-linear unconstrained optimization. Not surprisingly, the resulting solutions are identical to the parameters of the poisson model with a log link fuction if we permit z_M to be large enough.[1].

[1]In empirical work, it turns out that setting z_M to a value about twice as high as the largest value in the observed sample is sufficient to obtain parameter estimates identical to the Maximum Likelihood Poisson model

4. SOME EXTENDED FORMULATIONS

In this section we discuss various extensions to the basic model described above that address the various challenges that researchs typically face when analyzing real-world data.

4.1. Extra-Poisson Variation

A problem that is typically encountered in applied work with count outcomes is that of overdispersion. Poisson models have the restrictive assumption that the first two moments of a signal are identical. In most applied work, and for a host of reasons, this assumption can be violated and researchers have invented an impressive array of approaches to test for and account for this real-world feature. Basically, these solutions all boil down to the introduction of an overdispersing random variable inside the link function and various assumptions about this random variable (or its exponent) yield different models. For example, assuming that the exponent of this variable follows a gamma distribution yields the negative binomial model for count outcomes. Among the various Negative Binomial models, various parameterizations linking the variance of the overdispersed count outcome to its mean exist. The choice among these options is typically a matter of mathematical convenience and/or software availability. Reaserchers have also proposed finite mixture models that replace the assumption of a continuous random disturbance term with that of a discrete disturbance term with several (empirically determined) points of support.

In this paper, we take a slightly different approach. Noting that under the informatio-theoretic formulation, as described in the previous section, the introduction of the prior probability of $p_m^0 = \frac{1}{z_m!}$ was somewhat arbitrary, we proceed by questioning this restrictive assumption.

The solution for the Poisson regression model as obtained in (3.6) may be written in

a slightly different way, by taking the priors inside the exponent, as follows:

$$p_{mn} = \Omega_n^{-1} \exp(z_m \mathbf{x}_n' \boldsymbol{\lambda} - \log(z_m!)) \tag{4.1}$$

which immediately suggests that setting the priors to $\frac{1}{z_m!}$ results in a model that places a restriction of -1 on the term $\log(z_m!)$. Why not relax this restriction by parameterizing the dependence of the probability on the $\log(z_m!)$ term directly and by letting the data determine the magnitude of this dependence?

As noted by Zellner and Highfield (1988), Ryu (1993), and Golan, Judge, and Miller (1996), among others, Maximum (and Cross) Entropy distributions are solutions to particular kinds of constrained optimization problems where the constraints are suitable moment restrictions. Therefore, to obtain a solution that includes a parameter on the term $\log(z_m!)$ we need only introduce appropriately constructed moment constraints. It is easy to see that the moment we need to constrain in order to obtain this parameterization is $\sum_n \sum_m \log(z_m!) p_{mn}$ and the natural quantity to set this moment equal to is its sample analogue. Therefore, in addition to the moment constraints of the previous section, if we introduce the constraint

$$-\sum_n \log(y_n!) = -\sum_n \sum_m \log(z_m!) p_{mn}, \tag{4.2}$$

then the solution we obtain is

$$p_{mn} = \Omega_n^{-1} p_m^0 \exp(z_m \mathbf{x}_n' \boldsymbol{\lambda} - \delta \log(z_m!)) \tag{4.3}$$

where the new set of priors are assumed to be uniform (i.e., $p_m^0 = \frac{1}{M}\ \forall m$). Now, we may test the restriction underlying the Poisson model—that the parameter δ be set to 1. In effect, this

extension allows us to model overdispersion without assuming any specific overdispersing random variable but rather by relaxing a restrictive assumptions we had made to derive the Poisson model in the first place.

To see the similarity of the Cross Entropy problem (3.8) and a more flexible one with a restriction of $\delta = 1$, note that this *restricted* dual would be written as

$$\mathscr{L}_* = \sum_{kn} \lambda_k x_{kn} y_n - 1 \cdot \sum_n \log(y_n!) - \sum_n \Big[\sum_m p_m^0 \exp(z_m \mathbf{x}_n' \boldsymbol{\lambda} - 1 \cdot \log(z_m!)) \Big]. \tag{4.4}$$

Since addition of a term not involving any of the parameters with respect to which a function is being optimized does not alter the optimum solutions, the solutions resulting from optimizing (4.4) will be identical to those obtained by optimizing (3.8).

4.2. Systematic Extra-Poisson Variation

It is evident that the flexibility allowed in the last section can be futher extended to allow systematic variation in the δ parameter. That is, we may think of the added constraint as using only the first column of the design matrix to impose constraints on the observed values of $-\log(y_n!)$. A natural extention to this flexibility is to impose (and explicitly test) moment constraint utilizing covariance structures between the x_{kn} and $-\log(y_n!)$. To do that we can include, in the constrained optimization formulation, the following set of constraints

$$-\sum_n x_{kn} \log(y_n!) = -\sum_n x_{kn} \sum_m \log(z_m!) p_{mn} \qquad \forall k \tag{4.5}$$

so that the optimum solutions are now

$$p_{mn} = \Omega_n^{-1} p_m^0 \exp(z_m \mathbf{x}_n' \boldsymbol{\lambda} - \log(z_m!) \mathbf{x}_n' \boldsymbol{\delta}) \tag{4.6}$$

4.3. Preponderence of zero counts

In many applied setting the outcome variable has a pre-ponderence of 0 counts. There may be theoretical (or empirical) reason to expect that the mechanism resulting in some counts versus none is quantitatively or qualitatively different from the mechanism resulting in the number of counts. For example, in the literature on criminal activity, it is argued that what motivates people to commit some crime (versus none) may be different, qualitatively and quantitatively, from the behavioral model explaining the number of crimes they commit. The extension to allow for a preponderence of zero counts can be easily incorporated in the above framework by creating additional constraints on individual elements of the probability vector $\mathbf{p}_n$.

One can define, for example, a new binary choice outcome $y_{2n} = 1[y_n = 0]$ where $1[\cdot]$ is an indicator function yielding 1 if the condition $[\cdot]$ is satisfied and 0 otherwise. Since p_{1n} is the probability associated with the proposition $z_m = 0$, i.e., that no counts will be recorded, and since $\sum_{m=2}^{M} p_{mn}$ is therefore the probability that some counts will be observed, we can create a new set of constraints using the new outcome y_{2n}. Defining a new support space, $\mathbf{z}_2 = (1, 0, 0, \ldots, 0)'$ we can define the signal that results in the new outcome y_{2n} as $\mathbf{z}_2'\mathbf{p}_n$. Note that both sets of signals ($\mathbf{z}'\mathbf{p}_n$ and $\mathbf{z}_2'\mathbf{p}_n$) are constructed from the same set of probabilities and, therefore, the source of uncertainty in the model remains fixed. We are simply formulating an additional set of constraints on the same set of probabilities. These additional constraints can be written as

$$\sum_n x_{k_2 n} y_{2n} = \sum_n x_{k_2 n} \mathbf{z}_2 \mathbf{p}_n \qquad \forall k_2 = 1, \ldots, K_2 \tag{4.7}$$

so that the solution for the probabilities of interest are obtained as

$$p_{mn} = \Omega_n^{-1} p_m^0 \exp(z_m \mathbf{x}_n' \boldsymbol{\lambda} + z_{2m} \mathbf{x}_{2n}' \boldsymbol{\lambda}_2 - \delta \log(z_m!)) \qquad \forall m, n. \tag{4.8}$$

Note that, by definition of the problem, when $z_m = 0$ then $z_{2m} = 1$ and when $z_m > 0$ then $z_{2m} = 0$. Hence, there are two sets of lagrange multipliers operate over two different regions of the original support space $\mathbf{z}$. Moreover, the set of characteristics $x_{k_2 n}$ can be the same or different from the orginal set of covariates x_{kn}.

In addition to accomodating a pre-ponderence of 0 counts, nothing precludes researchers from specifying a preponderence of any other count. For example, when modeling the number of months (count of months) that defendants are senteced to prison for, judges typically impose sanctions at rounded-off months (eg., 12 months, 18 months, 60 months, etc.). Hence, there may be a preponderence of zero counts (when defendants are not sentenced to any prison) or a preponderence of other points in the support space. This setting may require specification of models that allow abnormally high counts at various other points on the support space.

4.4. Truncated counts and censored samples

In several real-world setting the data may be sampled from only a truncated part of the true realm of possibility. In order to deal with such truncated counts in the information-theoretic setting, it is trivial to restrict the support space to that region where we know the outcomes were sampled from. For example, if the outcomes were truncated between the counts of 0 and 9, then all we need do is define the support space as $\mathbf{z}_- = (0, 1, \ldots, 9)'$ and create moment constraints using these truncated signals. Another example where truncated count models would be applicable is if only positive counts were observed (e.g., surveys

of prisoners that may ascertain fro them for the number of crimes they have committed in the past). Clearly, any variation in the *observed* outcome is subject to this truncation and, therefore, we need only ensure that the variation we allow in the *expected* outcome (the signal) is subject to the same truncation. In other words, having defined the signal support as above, we can create moment constraints of the form

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}'_- \mathbf{p}_n \qquad \forall k \tag{4.9}$$

where the objective function is now defined only over probabilities corresponding to the truncated regions of the support space (i.e., $\mathbf{p}'_{-n} \log(\mathbf{p}_{-n}/\mathbf{p}^0_{-n})$). Once we recover the largange multipliers from this problem, we can then recover the complete untruncated signals by applying these lagrange multipliers to the complete untruncated region.[2]

In addition to the problem of truncated outcomes, sometimes researchers face censored samples whereby the outcomes for some portion of the sample are not observed or are incorrectly recorded. For example, it is not uncommon to see surveys that allow a single digit response to a question like "how many times did you experiece event E?" and "check this box if more then 9." This would indicate that for a part of the sample we have outcomes truncated in the region 0 through 9 and for the remaining part of the sample we only know that the count was above 9.

In addition to the fact that the signals from the uncensored part of the sample need to be truncated, we now have fewer outcomes observed than we have signals emitted. That is, from the total sample of N units, we observe an uncensored outcome for only $N_1 < N$ units. Hence we can set up two sets of constraints—one as described above for the case of truncated counts and another that models the probability of censoring. If censoring is at a

[2]The truncated setting discussed here applies only to *exogenous* truncation and not to *incedental* truncation.

certain point on the support space, then it is easy to define the probability of being selected in the uncensored sample as some function of all the probabilities corresponding to the truncated part of the support. Let us denote the probabilities over the truncated part of the support space by $\mathbf{p}_{-n}$. Let the fact that an observation is not censored be defined as a second outcome (say y_{2n}) and let us define a new support space $\mathbf{z}_2 = (1, 1, \ldots, 1, 0, 0, \ldots, 0)'$ that is set to 1 for all points that fall within the truncated region of the support space and 0 for all points that fall outside that truncated region. Then we have the following set of constraints

$$\sum_{n \in N_1} x_{kn} y_n = \sum_{n \in N_1} x_{kn} \mathbf{z}'_{-} \mathbf{p}_{-n} \qquad \forall k \tag{4.10}$$

$$\sum_{n} x_{k_2 n} y_{2n} = \sum_{n} x_{k_2 n} \mathbf{z}_2 \mathbf{p}_n \qquad \forall k_2 \tag{4.11}$$

where uncertainty is defined over the entire untruncated realm of possibility. We are now introducing the additional requirements that the process determining the count outcome is different from the process that determines the censoring outcome (while acknolwedging that the former can only be observed within a truncated region of the support space). Note that there is no reason for the two sets of constraints to have different sets of covariates—they are identified because the support spaces are distinct.

Once again, as wih the case of the truncated count model, once we have recovered the set of Lagrange multipliers we are after, we may then apply those to the entire realm of possible outcomes and recover complete untruncated signals for *all* units in the sample (censored and uncensored).

4.5. Endogenous switching

Unlike the sample selection case (above) where the binary endogenous choice is one of selection into and out of the truncated region, sometimes researchers are confronted with

the need to analyze count outcome models with binary choices as predictors where the binary choice may be endogenous. Rather than treat this as a selection process in and out of the truncated region of the support space, Terza (1998) uses the endogenous switching model that treats the binary endogenous choice as one of switching between two regimes. Under the Information-theoretic setting this means we only observe one outcome per observational unit. But, this outcome could be a result of a signal emitted from one of two independent sources. Let these signsl from the two distince sources be denoted by s_{1n} and s_{2n}. The probability that any given observation was a result of a signal emitted from one or the other of these two (mutually exclusive and exhaustive) sources may be defined as w_{1n} and w_{2n}. Then, if we define a new support space as $\mathbf{z}_2 = (0, 1)'$ and write $\mathbf{w}_n = (w_{1n}, w_{2n})'$ we obtain the following system of signal/outcome approximations.

$$y_n \approx w_{1n}\mathbf{z}'\mathbf{p}_{1n} + w_{2n}\mathbf{z}'\mathbf{p}_{2n} \qquad \forall n \tag{4.12}$$

$$y_{2n} \approx \mathbf{z}_2'\mathbf{w}_n \qquad \forall n \tag{4.13}$$

where $\mathbf{p}_{1n}$ and $\mathbf{p}_{2n}$ are now the conditional probabilities used to re-parameterize the signals form the two sources and y_{2n} is the endogenous binary choice.

We can now create two sets of moment constraints in order to convert the above inequalities to equalities. However, there are identifying restrictions that must be met. That is, the sets of covariates used in the two constraints may overlap but may not be identical. Also, since the model includes conditional probabilities, we will need to acknowledge this explicitly in the objective function when optimizing the KL directed divergence measure. The new objective function is defined as

$$\min_{\mathbf{p}_1,\mathbf{p}_2,\mathbf{w}} \quad \sum_{n,j=1,2} w_{jn}\log(w_{jn}/w_{jn}^0) + \sum_{n,j=1,2} w_{jn}\left\{\mathbf{p}_{jn}'\log(\mathbf{p}_{jn}/\mathbf{p}_{jn}^0)\right\} \tag{4.14}$$

In order to proceed we can define a composite probability measure $q_{jnm} = w_{jn} \cdot p_{jnm}$ such that w_{jn} and p_{jnm} are marginal and conditional probabilities of this measure. That is, we can write

$$1 = \sum_{jm} q_{jmn} \qquad \forall n \tag{4.15}$$

$$w_{jn} = \sum_{m} q_{jmn} \qquad \forall n, j = 1, 2 \tag{4.16}$$

$$p_{jmn} = \frac{q_{jmn}}{w_{jn}} = \frac{q_{jmn}}{\sum_m q_{jmn}} \qquad \forall n, m, j = 1, 2 \tag{4.17}$$

Using these definitions in the objective function, and letting $q^0_{mnj} = p^0_{mnj} \cdot w^0_{nj}$, we obtain a new objective function defined solely in terms of $\mathbf{q}$ as

$$\min_{\mathbf{q}} \quad \mathscr{K}(\mathbf{q} : \mathbf{q}^0) = \sum_n \sum_{j=1,2} \mathbf{q}'_{jn} \log(\mathbf{q}_{jn}/\mathbf{q}^0_{jn}) \tag{4.18}$$

and the new sets of constraints can be defined over the support spaces $\mathbf{z}_1 = (\mathbf{z}', \mathbf{z}')'$ and $\mathbf{z}_2 = (0, 0, \ldots, 0, 1, 1, \ldots, 1)'$ which are both $2M$ dimensional vectors of support spaces. These constraints can be written as

$$\sum_n x_{k_1 n} y_n = \sum_n x_{k_1 n} \mathbf{z}'_1 \mathbf{q}_n \tag{4.19}$$

$$\sum_n x_{k_2 n} y_{2n} = \sum_n x_{k_2 n} \mathbf{z}'_2 \mathbf{q}_n \tag{4.20}$$

with the adding up constraints $\mathbf{1}'\mathbf{q}_n = 1 \; \forall n$.

If there is reason to believe that some of the predictors have the same impact under the two regimes, then this would require setting the corresponding lagrange multipliers equal to zero in the regime selection constraints.

If the endogenous switching is between more than two regimes *and* we have knowl-

dge about the units that were observed in specific regimes, then we can model this in an analogous manner (to that defined above) with $J > 2$.

4.6. Nested Data Generating Structures

In several setting, researchers are confronted with mico-level units that are embedded within larger macro-level units. Two of the typical examples confronted in practice include Heirarchical models–where micro units are nested in one of several mutually exclusive macro units–and Repeated Measure models (also known as panel data sets)—where the same units are measured repeatedly over time thereby nesting temporal observations within individuals (persons, firms, etc.). In each of these setting, there may be a reason to be skeptical of any findings that ignores the nesting in the data because systematic variations are the macro level may be mistakenly attributed to the micro level simply because the structure was ignored. The basic idea in such setting seems to involve the recognition and incorporation of some unobserved commonality or *stickiness* among all units within a given macro unit. A simple way to proceed would be to introduce separate dummy variables capturing each of the macro units (less one). To do so under the information-theoretic approach would mean specifying some moment constraints only within a macro unit and some across macro units. Nothing precludes us from applying this approach under the current setting. However, there is the usual curse of dimensionality where the number of macro level intercepts may be very large (specially in panel data sets where we typically have large number of individuals followed for a finite number of repeated time periods). In this paper, we approach this propblem by directly allowing for within-macro-unit stickiness in the recovered signals.

In the previous sections, one of the implicit assumptions that was introduced in all the extensions discussed was that of $\sum_m p_{mn} = 1 \ \forall n$. In this section, we relax this constraint

somewhat to obtain the desired stickiness thereby re-formulating the information-recovery problem. Let us first write the generic multi-level signals and their manifestations as s_{nj} and y_{nj} respectively (where one may replace j by t to denote time). The approximation now becomes:

$$y_{nj} \approx s_{nj} \qquad \forall n = 1, \ldots, N_j; j = 1, \ldots, J \tag{4.21}$$

and with theory providing us with a set of K exogenous predictors for each of the micro-units, we could create simple moment constraints of the form

$$\sum_{nj} x_{knj} y_{nj} = \sum_{nj} x_{knj} \sum_{m} z_m p_{mnj} \qquad \forall k \tag{4.22}$$

along with any of the extended constraints discussed above. Additionally, the objective function to be optimized in the simple setting would be

$$\max_{\mathbf{p}} \quad \mathscr{K}(\mathbf{p}; \mathbf{p}^0) = \sum_{nj} \mathbf{p}'_{nj} \log \left(\mathbf{p}_{nj} / \mathbf{p}^0_{nj} \right) \tag{4.23}$$

The last set of constraints we would need to impose in order to complete the problem would be the adding up constraints on the probabilities of interest. That is, $\sum_m p_{mnj} = 1$ $\forall n, j$. However, if impose that constraint we explicitly rule out the stickiness we desire. In order to introduce the possibility of something unobserved but common to all micro-units within a macro-units we need to impose less restrictive constraints. Consider the less restrictive adding-up constraints

$$\sum_{mn} p_{mnj} = N_j \qquad \forall j \tag{4.24}$$

These constraints are implied by the more restrictive $\sum_m p_{mnj} = 1$ but not the converese. In other words, by imposing the adding up constraints as above, we would be reducing the

total number of constraints in the primal optimization problem. The problem we face with these less restrictive constraints is that they do not bind the probabilities p_{mnj} to be proper (i.e., $\in (0, 1)$). Therefore, we first define an auxiliary probability measure $N_j q_{mnj} = p_{mnj}$ so that we obtain a new adding up constraint of the form $\sum_{mn} q_{mnj} = 1 \; \forall j$. In addition, we can create comparable prior probabilities be setting $q^0_{mnj} = (1/N_j)p^0_{mnj}$. Finally, we replace p_{mnj} in the entire primal problem with $q_{mnj}N_j$. This yields the following primal Lagrange function

$$\begin{aligned} \mathscr{L} \;=\; & \sum_j \sum_{mn} N_j q_{mnj} \log\left(q_{mnj}/q^0_{mnj}\right) + \sum_j \eta_j \left\{1 - \sum_{mn} q_{mnj}\right\} \\ & \sum_k \lambda_k \left\{ \sum_{nj} x_{knj} y_{nj} - \sum_{nj} x_{knj} \sum_m z_m N_j q_{mnj} \right\} \end{aligned} \tag{4.25}$$

Solving the first order condition of this optimization problem, we obtain the solutions

$$q_{mnj} = \frac{q^0_{mnj} \exp(z_m \mathbf{x}'_{nj} \boldsymbol{\lambda})}{\sum_{mn} q^0_{mnj} \exp(z_m \mathbf{x}'_{nj} \boldsymbol{\lambda})} = \frac{q^0_{mnj} \exp(z_m \mathbf{x}'_{nj} \boldsymbol{\lambda})}{\Omega_j} \tag{4.26}$$

and, inserting this optimum solution into the primal problem of (4.25), we obtain the corresponding unconstrained dual objective as a function of the Lagrange multipliers

$$\mathscr{L}_* = \sum_{knj} x_{knj} y_{nj} \lambda_k - \sum_j N_j \log \Omega_j \tag{4.27}$$

Note that, given the assymetry in the implication of the adding up constraints, i.e., that $\sum_m p_{mnj} = 1 \; \forall n, j$ implies $\sum_{mn} p_{mnj} - N_j \; \forall j$ but not the converse, we cannot recover our probabilities of interest (p_{mnj}) by scaling the estimated q_{nmj} by the factor N_j. We must compute these as conditional probabilities. That is, we can recover the probabilities p_{mnj}

by

$$p_{mnj} = \frac{q_{mnj}}{\sum_m q_{mnj}} = \frac{p_m^0 \exp(z_m \mathbf{x}'_{nj} \boldsymbol{\lambda})}{\sum_m p_m^0 \exp(z_m \mathbf{x}'_{nj} \boldsymbol{\lambda})} \tag{4.28}$$

5. ASSESSING THE SAMPLING VARIABILITY OF THE LAGRANGE MULTIPLIERS

If the information-theoretic framework is employed for recovering a set of Lagrange multiplers pertaining to a particular problem from *one* specific sample, then it is of obvious interest to study how these recovered Lagrange multipliers may vary across different samples, i.e., to study their sampling variability. In this section we provide a brief discussion of the sensitivity of the Lagrange multipliers to sampling variability.

To keep the discussion and derivations below as generic as possible, we re-write the Cross Entropy dual objective function (3.8) as

$$\mathscr{L}_* = \sum_k \lambda_k \mu_k - f_s(\boldsymbol{\lambda}) \tag{5.1}$$

where $\mu_k = \sum_n x_{kn} y_n$ are the sample statistics. The Poisson dual objective functions may be obtained by appropriately specifying $f_s = \sum_n \ln \Omega_n$.

The optimal solutions for this unconstrained maximization problem is found by simultaneously solving the K first order conditions

$$\frac{\partial \mathscr{L}_*}{\partial \lambda_k} = \mu_k - \frac{\partial f_s(\boldsymbol{\lambda})}{\partial \lambda_k} = 0 \qquad \forall k \tag{5.2}$$

and ensuring that, at the optimal solutions, the Hessian matrix, computed as

$$\frac{\partial^2 \mathscr{L}_*}{\partial \hat{\lambda}_k \partial \hat{\lambda}_{k'}} = -\frac{\partial^2 f_s(\boldsymbol{\lambda})}{\partial \lambda_k \partial \lambda_{k'}} \qquad \forall k, k', \tag{5.3}$$

is negative definite. Given the logrithmic forms of the function f_s, the dual objective function is strictly concave thereby ensuring a unique global maximum if a maximum exists.

We can study variations in the optimal Lagrange multipliers that can be expected due to fluctuations in the sample statistics μ_k (from one sample to another) by taking the total derivative of the K first order conditions (5.2) with respect to each $\{\mu_k\}$ and $\{\hat{\lambda}_k\}$. In matrix notation, this may be written as the following system of K differential equations

$$\mathrm{d}\boldsymbol{\mu} - \frac{\partial^2 f_s(\hat{\boldsymbol{\lambda}})}{\partial\hat{\boldsymbol{\lambda}}\partial\hat{\boldsymbol{\lambda}}'}\mathrm{d}\hat{\boldsymbol{\lambda}} = \mathbf{0}. \tag{5.4}$$

Using the definition of the Hessian from (5.3) and rearranging terms we obtain the desired relationship between variations in the optimal Lagrange multipliers and variations in the sample statistics as

$$\frac{\mathrm{d}\hat{\boldsymbol{\lambda}}}{\mathrm{d}\boldsymbol{\mu}'} = \left\{ - \frac{\partial^2 \mathscr{L}_*}{\partial\hat{\boldsymbol{\lambda}}\partial\hat{\boldsymbol{\lambda}}'} \right\}^{-1} \tag{5.5}$$

This relationship implies that if we can make certain assumptions about how the sample statistics (μ_k) vary across repeated samples then we can make claims about the implied distribution of the Lagrange mulipliers across these samples. For the former, we can rely on the Central Limit Theorem according to which, irrespective of the population distribution of a random variable, computed sample statistics (such as sums or means) of this random variable taken across several samples of a given size will be normaly distributed even if these samples are as small as 30 to 40 units. Here the μ_k are one such sample statistic. Consequently, we may assume that the optimal Lagrange multipliers are normally distributed as well with an asymptotic covariance given by the RHS of (5.5)

6. SOME EMPIRICAL APPLICATION

(TO COME)