*Econometrics* **2015**, *3*, 187-198; doi:10.3390/econometrics3020187

OPEN ACCESS

*econometrics*

**ISSN 2225-1146**

www.mdpi.com/journal/econometrics

*Article*

# Information Recovery in a Dynamic Statistical Markov Model

**Douglas J. Miller [1] and George Judge [2,*]**

[1] Economics and Management of Agrobiotechnology Center, University of Missouri, Columbia, MO 65211, USA; E-Mail: doug@new-analytics.com

[2] Graduate School, 207 Giannini Hall, University of California, Berkeley, Berkeley, CA 94720, USA

* Author to whom correspondence should be addressed; E-Mail: gjudge@berkeley.edu; Tel.: +1-510-642-0791.

Academic Editor: Kerry Patterson

*Received: 17 October 2014 / Accepted: 25 February 2015 / Published: 25 March 2015*

---

**Abstract:** Although economic processes and systems are in general simple in nature, the underlying dynamics are complicated and seldom understood. Recognizing this, in this paper we use a nonstationary-conditional Markov process model of observed aggregate data to learn about and recover causal influence information associated with the underlying dynamic micro-behavior. Estimating equations are used as a link to the data and to model the dynamic conditional Markov process. To recover the unknown transition probabilities, we use an information theoretic approach to model the data and derive a new class of conditional Markov models. A quadratic loss function is used as a basis for selecting the optimal member from the family of possible likelihood-entropy functional(s). The asymptotic properties of the resulting estimators are demonstrated, and a range of potential applications is discussed.

**Keywords:** conditional moment equations; controlled stochastic process; first-order Markov process; Cressie-Read power divergence criterion; quadratic loss; adaptive behavior

**JEL Classification:** C40, C51

---

## 1. Introduction

In this paper we recognize that understanding and predicting the future state of an economic-behavioral process is statistical in nature and that the underlying dynamics may be complex and not well understood. Within this context, in order to develop a general quantitative approach to information recovery, we use a first order Markov process in discrete time and an information theoretic probability model to characterize, in a time ordered state space form, causal influence information in dynamic economic-data. To seek new ways to think about adaptive intelligent behavior of dynamic micro systems, we follow Wissner-Gross and Freer [1] and use causal entropy maximization or optimizing criterion as the systems status measure. While the emphasis is on recovering the unknown transition probabilities associated with the underlying micro process, the observed dynamic data often only exist in macro-aggregate form. Thus, information recovery involves the solution of an ill-posed inverse problem, and by its nature suggests the use of the tools of information theory.

In the context of a stochastic dynamic Markov economic process and a sample of time ordered aggregate data, we use a $k$, $t$ state-space Markov model, as a basis for recovering the unknown transition probability parameters $p(j,k,t)$ and the unknown underlying sample probability distribution function. In terms of a solution framework we recognize that although the transition parameters $p(j,k,t)$ are unobserved and unobservable, we usually do know some linear functionals or moments, $u(p(j,k,t))$, that are linked to the data—see for example [2,3]. Given the unknown underlying data distribution function, under this formulation there are many possible optimal solutions for $p(j,k,t)$, and the question arises as to how to choose among them. As a way of reducing the solution uncertainty, we use an information measure $I(p)$ to characterize a given probability distribution, and divergence measures to identify the distance between the candidate probability distributions. This permits us to exploit the statistical machinery of information theory to gain insights into the unknown transition parameters and the underlying probability distribution behavior of the dynamic state space process. As a solution basis we use the Cressie-Read [4] family of entropy-likelihood functionals, as a way to recover an estimator of the transition probabilities $p(j,k,t)$, and the underlying model error data process. In terms of estimator choice, we use in a statistical loss function context, a convex combination of entropy functionals-likelihoods from the Cressie-Read (CR) family.

## 2. The Markov Econometric Model and the Information Recovery Process

In this paper, we focus on first-order Markov chain models of events with a finite number of outcomes measured at discrete time intervals. From the micro perspective, the decision outcomes for agent $i = 1,2,\ldots,n$ are denoted $Y(i,k,t)$ with finite states $k = 1,2,\ldots,K$ at time $t = 0,1,2,\ldots T$, where

$$Y(i,k,t) = \begin{cases} 1 \text{ if agent } i \text{ selects state } k \text{ at time } t \\ 0 \quad otherwise \end{cases}$$

In this context, we consider a finite economic process that involves the outcome space in the form of agents that are driven by a maximization principle. If the decision outcomes exhibit first-order Markov character, the dynamic behavior of the agents may be represented by conditional transition probabilities $p(j,k,t)$, which represent the probability that agent $i$ moves from state $j = 1,2,\ldots,K$ to state $k$ at a time $t$. Given observations on the micro behavior $Y(i,k,t)$, a number of researchers have used the

discrete Markov decision process framework to model the agent-specific dynamic economic behavior (see for example [5–7]).

In practice, the observed information may be limited to the aggregated outcomes, $Y(k,t) = n^{-1}\sum_{i=1}^{n} Y(i,k,t)$ for each $k$ and $t$, and the approach to estimating the Markov transition probabilities, is based on the estimating equations

$$Y(k,t) = \sum_{j=1}^{K} Y(j,t-1)p(j,k,t) \tag{2.1}$$

that link the observed aggregate data to the transition probabilities. The transition probabilities may be directly estimated from Equation (2.1) if the Markov process is unconditional or stationary such that $p(j,k,t) = p(j,k,t+s)$ for all integers $s$. In this case, the estimation problem reduces to a set of $(K-1) \times T$ estimating equations with $(K-1) \times (K-1)$ unknown Markov transition probabilities, $p(j,k)$, which fulfill the row-sum conditions, $\sum_{k=1}^{K} p(j,k,t) = 1$. For this case, Lee, Judge, and Zellner [8] use least squares, quadratic programming, or other established estimation procedures to directly compute estimates of the unconditional Markov transition probabilities if $T \geq K$.

### *2.1. Sample Analogs of the Markov Process*

In this paper we are concerned with nonstationary or conditional Markov models in which $p(j,k,t)$ varies with $t$. In the conditional case, we have $T \times (K-1) \times (K-1)$ unknown transition probabilities in the $(K-1) \times T$ estimating equations, and the estimation problem is ill-posed. One way to overcome the ill-posed nature of the problem is to specify parametric functional forms for the Markov transition probabilities. For example, MacRae [9] and Theil [10] recommend the logistic functional form

$$p(j,k,t) = \frac{\exp(\beta_{jk}' z_{t-1})}{\sum_{k=1}^{K} \exp(\beta_{jk}' z_{t-1})} \tag{2.2}$$

where $z_{t-1}$ is a vector of explanatory-instrumental variables and $\beta_{jk}$ is a vector of conformable state-specific and time-invariant model parameters. The parametric logistic specification has some advantages for the purposes of estimation and interpretation since the transition probabilities satisfy $p(j,k,t) > 0$ and the row-sum conditions.

Since for estimation and inference, the available data may be partial or incomplete, one key step is linking the Markov process to the indirect noisy observations. In this context we note that the sample analog is a noisy version of the traditional estimating Equation (2.1). Accordingly, the sample of analog Equation (2.1) may be stated as

$$Y(k,t) = \sum_{j=1}^{K} Y(j,t-1)p(j,k,t) + e(k,t) \tag{2.3}$$

Following Miller [2], the error term $e(k,t)$ may be formulated as a first-order vector moving average (VMA) process in the underlying agent-specific innovations. Under modest assumptions on the agents' Markov decision process, this composite error term has null mean and may be conditionally heteroskedastic. We discuss the implications of these properties of $e(k,t)$ for the large sample properties

of the proposed parameter estimators in Section 5. For empirical purposes, the remaining task is to choose a feasible specification of the statistical model of the Markov transition probabilities.

*2.2. Modeling the Conditional Transition Probabilities*

In the next section we derive a new class of conditional Markov models. The proposed model of the process is based on a set of estimating equations-moment equations

$$\mathrm{E}[z_t' e(k,t)] = 0 \tag{2.4}$$

Where $\mathbf{z}_t$ is an appropriate set of instrumental-intervention variables and the model error is assumed to be ergodic such that the expected value is suitably defined. By substitution of Equation (2.3) into (2.4), we form a set of estimating equations that are expressed in terms of the unknown transition probabilities.

$$E\left[z'_t\left(Y(k,t) - \sum\nolimits_{j=1}^{K} Y(j,t-1)p(j,k,t)\right)\right] = 0 \tag{2.5}$$

This set of moment-estimating equations depicts the Markov process in terms of exogenous and endogenous variables. The endogenous state transition probabilities are conditioned on the exogenous variables and produce a conditional Markov process. Given there may be many feasible transition probability models that satisfy the moment-estimating equations, the next step is to provide a model for the data [11], and a basis for identifying parametric data sampling distributions and likelihood functions in the form of distances in probability space. In this context, the Cressie-Read (CR) family of divergence measures provides access to a rich set of distribution functions that permits a basis for identifying the parametric data sampling distribution. This family of divergence measures permits us to recognize that the natural solution may not be a fixed distribution, but a well-defined set of distributions-likelihood functions. The CR flexible family of divergence measures is introduced in Section 3 and permits us to seek, given a sample of data, the optimum convex combination of likelihood functions, under a quadratic loss measure. This leads to estimators of shrinkage form and permits us to derive a conditional Markov model from the sample analogs of Equations (2.3) and (2.4).

## 3. Cressie-Read Power Divergence (PD) Criterion

The Cressie-Read (CR) power divergence (PD) statistic [4,12,13], may be defined for a set of first-order finite and discrete conditional Markov probabilities as

$$I\,(p,q,\alpha) = \frac{1}{\alpha\,(1+\alpha)}\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} p(j,k,t)\left[\left(\frac{p(j,k,t)}{q(j,k,t)}\right)^{\alpha} - 1\right] \tag{3.1}$$

Here, the PD statistic measures the pseudo-distance between $p$ and a set of reference transition probabilities $q$. We note that Equation (3.1) encompasses a family of estimation objective functions indexed by discrete probability distributions. For our purposes, the PD statistic is a useful estimation criterion due to its information theoretic properties such as strict convexity in $p$.

The CR family of entropic-likelihood functionals and the corresponding convex combinations of its family members' represent a very useful collection of divergences (see [14,15]). In order not to

introduce subjective information, the reference distribution in the formulations to come will be specified as discrete uniform distributions.

*Minimum Power Divergence (MPD) Models*

Given reference weights $q$ and the orthogonality conditions Equation (2.4), our proposed method for deriving a model of the conditional Markov transition probabilities is to choose $p$ that satisfies the estimating equations and is closest to $q$ under the PD criterion. The resulting minimum power divergence (MPD) transition probabilities satisfy the behavioral conditions of the Markov decision process while remaining "least-informative" relative to the set of reference weights. Formally, the MPD problem may be solved by choosing transition probabilities $p$ to minimize $I(p,q,\alpha)$ (for some α) subject to the sample analogs of Equations (2.3) and (2.4)

$$\sum_{t=1}^{T} \boldsymbol{z}_t' \left( Y(k,t) - \sum_{j=1}^{K} Y(j,t-1)p(j,k,t) \right) = \mathbf{0} \tag{3.2}$$

for each $j = 2, \ldots, K$ and the row-sum constraint

$$\sum_{k=1}^{K} p\,(j,k,t) = 1 \tag{3.3}$$

for all $j$ and $t$.

If, for expository purposes, we consider cases where the reference weights are discrete uniform, when $\alpha \to 0$ the entropy-likelihood functional is

$$-I(p,q,\alpha \to 0) \propto -\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} p(j,k,t)\ln\big(p(j,k,t)\big) \tag{3.4}$$

When $\alpha \to -1$, the entropy-empirical likelihood functional [16] is

$$-I(p,q,\alpha \to -1) \propto \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} \ln\big(p(j,k,t)\big). \tag{3.5}$$

and the Maximum Likelihood objective function is defined. Correspondingly, the Kullback-Leibler directed divergence or discrimination information statistic with non-uniform reference weights $q(j,k,t)$ [17,18], is

$$I(p,q,\alpha \to 0) \propto \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} p(j,k,t)\ln\left(\frac{p(j,k,t)}{q(j,k,t)}\right)$$

and

$$I(p,q,\alpha \to -1) \propto \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} q(j,k,t)\ln\left(\frac{q(j,k,t)}{p(j,k,t)}\right) \tag{3.6}$$

In the case of $\alpha \to 0$ for Equation (3.4), the MPD solution produces Markov transition probabilities that have a general logistic functional form but different arguments than Equation (2.2)

$$\hat{p}(j,k,t) = \frac{\exp\left(-Y(j,t-1)z_t'\hat{\boldsymbol{\lambda}}_k\right)}{\sum_{k=1}^{K}\exp\left(-Y(j,t-1)z_t'\hat{\boldsymbol{\lambda}}_k\right)} \tag{3.7}$$

under uniform reference weights. In general, the MPD solution with non-uniform reference weights takes the form

$$\tilde{p}(j,k,t) = \frac{q(j,k,t)\exp\left(Y(j,t-1)z_t'\widetilde{\lambda_K}\right)}{\sum_{k=1}^{K}q(j,k,t)\exp\left(Y(j,t-1)z_t'\widetilde{\lambda_K}\right)} \tag{3.8}$$

The vectors $\tilde{\lambda}_K$ and $\tilde{\lambda}_K$ are the set of optimal Lagrange multipliers for the constraints Equation (3.2) under MPD objectives Equations (3.4) and (3.6), respectively. A closed-form solution for $\tilde{\lambda}_K$ and $\tilde{\lambda}_K$ do not exist in general, and estimates of $p(j,k,t)$ must be numerically determined. This entropy-based approach yields a model of the Markov transition probabilities in terms of the observed outcome variables and a finite number of estimable transition probabilities or parameters.

Under the usual regularity conditions on the moment-estimating equations, the large-sample properties of consistency, asymptotic normality and efficiency of the MPD estimator follow. Even if the model is not correctly specified, the MPD estimator may converge in probability and be asymptotically normal. Under these large sample results, classical hypothesis tests may apply for the MPD Markov formulation. The sampling properties of the MPD estimator are developed in Section 5 of this paper.

## 4. The Optimal MPD Estimator Choice under KL and Quadratic Loss

The likelihood function and the sample space for the MPD estimation problem are inexplicably linked and it would be useful, given a sample of indirect noisy observations and corresponding moment conditions, to have an optimum choice of a member of the CR family. However, at this point we should recognize that the solution may not be a fixed distribution, but a well-defined set of distributions-likelihood functions-PDFs.

### *4.1. Distance-Divergence Measures*

In Section 3, we used the CR power divergence measure,

$$I\,(p,q,\alpha) = \frac{2}{\alpha\,(1+\alpha)}\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}p(j,k,t)\left[\left(\frac{p(j,k,t)}{q(j,k,t)}\right)^{\alpha}-1\right] \tag{4.1}$$

to define a family of likelihood functions. Given this family of likelihood functions, one might follow Gorban [19] and Gorban and Karlin [20], and consider a parametric family of concave entropy-likelihood functions, which satisfy additivity and trace conditions. Using the CR divergence measures, this parametric family is essentially the linear convex combination of the cases where $\alpha \to 0$ and $\alpha \to -1$. This family is analytically tractable and provides a basis for joining (combining) statistically independent subsystems. When the base measure of the reference distribution $q(j,k,t)$ is taken to be a uniform probability mass function (PMF), we arrive at a one-parameter family of additive convex functions. From the standpoint of extremum-minimization with respect to the transition probabilities, the generalized divergence family which in the case of uniform $q(j,k,t)$ reduces to

$$S_{\beta}^{*} = -(1-\beta)\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} p(j,k,t)\ln\big(p(j,k,t)\big) + \beta\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\ln\big(p(j,k,t)\big) \tag{4.2}$$

which is a convex combination of the CR family members in Equations (3.3) and (3.4) with weight $0 \leq \beta \leq 1$.

In the limit, as $\beta \to 0$, the Kullback-Leibler or minimum I-divergence $I(p,q,\alpha \to 0)$ is recovered. As $\beta \to 1$, the MEL stochastic inverse problem $I(p,q,\alpha \to 0)$ results. *This generalized family of divergence measures permits a broadening of the canonical distribution functions and provides a framework for developing a loss-minimizing estimation rule.* In an extremum estimation context, when $\beta = 1/2$, the resulting estimation criterion is commonly known in the literature as Jeffrey's J-divergence. In line with the complex nature of the problem, in the sections to follow, we demonstrate alternative convex estimation rules, which seek to choose among MPD-type estimators to minimize the quadratic risk (QR) objective in Equation (4.2).

### *4.2. A Minimum Quadratic Risk (QR) Estimation Rule*

In this section, we use the well-known squared error-quadratic loss criterion and associated QR function to make optimal use of a given set of discrete alternatives for the CR goodness-of-fit measures and associated estimators for the transition probabilities (see for example, [14]). The method seeks to define the convex combination of a set of estimators that minimize QR, where each estimator is defined by the solution to the extremum problem Equation (4.1) subject to Equations (3.2) and (3.3)

$$\widehat{\mathbf{p}}(\alpha) = \arg\max_{\mathbf{p}}\{-I(p,q,\alpha) \,|\, \textstyle\sum_t \mathbf{z}_t' e(k,t) = 0, \sum_k p(j,k,t) = 1, p(j,k,t) \geq 0\} \tag{4.3}$$

where $\hat{p}(\alpha)$ is the vectorized set of transition probability estimators for a given α.

The squared error loss function for this problem is defined by $\boldsymbol{\rho}(\hat{\mathbf{p}},\mathbf{p}) = (\hat{\mathbf{p}}-\mathbf{p})'(\hat{\mathbf{p}}-\mathbf{p})$ and has the corresponding QR function given by

$$\boldsymbol{\rho}(\hat{\mathbf{p}},\mathbf{p}) = \mathrm{E}[\boldsymbol{\ell}(\hat{\mathbf{p}},\mathbf{p})] = \mathrm{E}[(\hat{\mathbf{p}}-\mathbf{p})'(\hat{\mathbf{p}}-\mathbf{p})]. \tag{4.4}$$

The convex combination of estimators is defined by

$$\bar{\mathbf{p}}(\boldsymbol{\beta}) = \sum_{h=1}^{H} \beta_h \, \widehat{\mathbf{p}}(\alpha_h)$$

where

$$\beta_h \geq 0 \text{ and } \textstyle\sum_{h=1}^{H} \beta_h = 1 \tag{4.5}$$

The optimum use of the discrete alternatives under QR is determined by choosing the particular convex combination of the estimators that minimizes QR, as

$$\bar{\mathbf{p}}(\hat{\beta}) = \textstyle\sum_{h=1}^{H} \widehat{\beta_h}\widehat{\mathbf{p}}(\alpha_h) \text{ where } \widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in CH}\{\rho(\bar{\mathbf{p}}(\boldsymbol{\beta}),\mathbf{p})\} \tag{4.6}$$

and *CH* denotes the *H*-dimensional convex hull of possibilities for the $\boldsymbol{\beta}$ vector, defined by the nonnegativity and adding-up conditions represented in Equation (4.5).

This represents one possible method of addressing the combining issue, relative to an appropriate choice of the gamma parameter, in the definition of the CR power divergence criterion.

*4.3. The Case of Two Alternatives*

In this section, we consider the case where there are two discrete alternative CR measures of interest. In this context, the analyst wishes to make optimal use of the information contained in the two associated estimators of $\mathbf{p}, \hat{\mathbf{p}}(\alpha_1)$ and $\mathbf{p}(\alpha_2)$. The corresponding QR function may be written as

$$\rho(\bar{\mathbf{p}}(\beta),\mathbf{p}) = E\big[[\beta(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p}) + (1-\beta)(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})]'[\beta(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p}) + (1-\beta)(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})]\big] \tag{4.7}$$

and can be represented in terms of the QR functions of $\hat{\mathbf{p}}(\alpha_1)$ and $\hat{\mathbf{p}}(\alpha_2)$ as

$$\rho(\bar{\mathbf{p}}(\beta),\mathbf{p}) = \beta^2\rho(\hat{\mathbf{p}}(\alpha_1),\mathbf{p}) + (1-\beta)^2\rho(\hat{\mathbf{p}}(\alpha_2),\mathbf{p}) + 2\beta(1-\beta)E[(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p})'(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})] \tag{4.8}$$

To minimize $\rho(\bar{\mathbf{p}}(\beta), \mathbf{p})$, the first-order condition, with respect to $\beta$, is given by

$$\frac{d\rho(\bar{\mathbf{p}}(\beta),\mathbf{p})}{d\beta} = 2\beta\rho(\hat{\mathbf{p}}(\alpha_1),\mathbf{p}) - 2(1-\beta)\rho(\hat{\mathbf{p}}(\alpha_2),\mathbf{p}) + 2(1-2\beta)E[(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p})'(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})] = 0 \tag{4.9}$$

Solving for the optimal value of $\beta$ yields

$$\hat{\beta} = \frac{\rho(\hat{\mathbf{p}}(\alpha_2),\mathbf{p}) - E[(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p})'(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})]}{\rho(\hat{\mathbf{p}}(\alpha_1),\mathbf{p}) + \rho(\hat{\mathbf{p}}(\alpha_2),\mathbf{p}) - 2E[(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p})'(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})]} \tag{4.10}$$

and the optimal convex-combined estimator is defined as

$$\bar{\mathbf{p}}(\hat{\beta}) = \hat{\beta}\hat{\mathbf{p}}(\alpha_1) + (1-\hat{\beta})\hat{\mathbf{p}}(\alpha_2) \tag{4.11}$$

By construction, $\bar{\mathbf{p}}(\hat{\beta})$ is QR superior to either $\hat{\mathbf{p}}(\alpha_1)$ or $\hat{\mathbf{p}}(\alpha_2)$, unless the optimal convex combination resides at one of the boundaries for $\beta$, or the two estimators have identical risks and $E[(\hat{\mathbf{p}}(\alpha_1) - \mathbf{p})'(\hat{\mathbf{p}}(\alpha_2) - \mathbf{p})] = 0$. In any case, QR wise, the resulting estimator $\bar{\mathbf{p}}(\hat{\beta})$ is certainly no worse, QR-wise, than either $\hat{\mathbf{p}}(\alpha_1)$ or $\hat{\mathbf{p}}(\alpha_2)$.

## 5. Sampling Properties of the MPD Estimators

To demonstrate the large-sample properties of the MPD parameter estimators, we consider the special case of the Kullback-Leibler cross-entropy functional (*i.e.*, $\alpha \to 0$). First, the constrained MPD problem may be reduced to an unconstrained form by concentrating the Lagrangian objective function. If we view the implicitly defined Markov transition probabilities Equations (3.7) and (3.8) as intermediate solutions and substitute these forms back into the Lagrangian expression, we can form a concentrated objective function that only depends on the vector of Lagrange multipliers, $\lambda$. In particular, the concentrated MPD objective function associated with Equation (3.8) reduces to

$$m(\lambda) = \sum_{t=1}^{T}\sum_{k=1}^{K} Y(k,t)z_t'\lambda_k - \sum_{t=1}^{T}\sum_{j=1}^{K} \ln\left[q(j,1,t) + \sum_{k=2}^{K} q(j,k,t)\exp(Y(j,t-1)z_t'\lambda_k)\right] \tag{5.1}$$

By the saddle-point property of the constrained minimization problem, Equation (5.1) is strictly concave in $\boldsymbol{\lambda}$ such that $m(\lambda) < m(\tilde{\lambda})\ \forall\ \lambda \neq \tilde{\lambda}$. Accordingly, the optimal values of $\boldsymbol{\lambda}$ can be computed

by maximizing $m(\lambda)$ by choice of $\boldsymbol{\lambda}$. The gradient vector of $m(\lambda)$ is simply the set of sample analog estimating Equation (3.2), and Newton-Raphson or related optimizations algorithms may be used to compute the optimal Lagrange multipliers.

The large-sample properties of the MPD estimator (including the minimum cross-entropy and maximum entropy estimators as special cases) may be derived under the following regularity conditions:

- **B1**: There exists $\lambda^0$ such that $p(\lambda^0; j, k, t) = p^0(j, k, t)$ for all $j$, $k$, and $t$.
- **B2**: The sample analog Equation (3.2) is consistent such that

$$T^{-1}\sum_{t=1}^{T} z'_t \left( Y(k,t) - \sum\nolimits_{j=1}^{K} Y(j,t-1)p^0(j,k,t) \right) \xrightarrow{p} 0 \tag{5.2}$$

- **B3**: The moment condition Equation (3.2) is asymptotically normal as

$$T^{-1/2}\sum_{t=1}^{T} z'_t \left( Y(k,t) - \sum\nolimits_{j=1}^{K} Y(j,t-1)p^0(j,k,t) \right) \xrightarrow{d} N(0,\Delta) \tag{5.3}$$

Note that we should expect the asymptotic covariance matrix to exhibit heteroskedastic and autocorrelated (HAC) character due to the properties of the composite model errors. Following our discussion of Equation (2.3), $e(k,t)$ is serially correlated because it is composed as a VMA process in the agent-specific innovations in the underlying Markov decision process. Further, the weights in the moving averages are based on the Markov transition probabilities, so the variance of $e(k,t)$ may be conditionally heteroskedastic if the Markov transition probabilities $p(j,k,t)$ are nonstationary and vary with $t$. Given these conditions, the consistency and asymptotic normality of the MPD estimator follows under:

- **Proposition 1**: The MPD estimator is consistent such that $\tilde{\lambda} \xrightarrow{p} \lambda_0$ under Assumptions B1 and B2 plus:
    1. there exists function $m_0(\lambda)$ that is uniquely maximized at $\lambda^0$
    2. $m(\lambda)$ is twice continuously differentiable and concave
    3. $T^{-1}m(\lambda) \xrightarrow{p} m_0(\lambda)$ for all $\boldsymbol{\lambda}$
- **Proposition 2**: The MPD estimator is asymptotically normal as

$$\sqrt{T}\left(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Gamma}_0^{-1}\Delta\boldsymbol{\Gamma}_0^{-1}) \tag{5.4}$$

- under the conditions of Proposition 1 plus Assumption B3 and
    1. there exists $\Gamma(\lambda)$ continuous in $\boldsymbol{\lambda}$ such that

$$\sup_{\lambda} \left\| T^{-1}\frac{\partial^2 m(\lambda)}{\partial\lambda\,\partial\lambda'} - \Gamma(\lambda) \right\| \xrightarrow{p} 0 \tag{5.5}$$

    2. $\Gamma_0 = \Gamma(\lambda_0)$ is nonsingular

In particular, the stated regularity conditions may be used to prove Propositions 1 and 2 under Theorems 2.7 and 3.1 provided by Newey and McFadden [21].

Following the test results developed for the frequency-based data cases by Kelton and Kelton [22], we can use the large-sample results presented above to conduct classical hypothesis tests under the

fitted MPD Markov models. In particular, we can test sets of linear restrictions under the general null hypothesis $H_0 : \mathbf{c}\lambda = \mathbf{r}$ with the Wald statistic

$$W = T\left(\mathbf{c}\tilde{\lambda} - \mathbf{r}\right)'\left(\widetilde{\Gamma^{-1}}\tilde{\Delta}\widetilde{\Gamma^{-1}}\right)^{-1}\left(\mathbf{c}\tilde{\lambda} - \mathbf{r}\right) : \overset{a}{\rightarrow} \chi_Q^2 \tag{5.6}$$

where matrix **c** has full row rank of $Q$. For example, we can conduct a Wald test for time-invariance (*i.e.*, stationarity) of the Markov process under the null hypothesis $H_0 : \lambda^* = 0$ where $\lambda^*$ is the subset of the Lagrange multipliers associated with non-constant elements of $\mathbf{z}_t$ and the reference weights $q(j,k,t)$ do not vary with $t$. Following the discussion above, the Wald statistic should incorporate an HAC-consistent covariance estimator (e.g., [23]) in order to accommodate the HAC structure of the underlying error process.

At this point, it is important to note that the proposed estimators are also asymptotically efficient if the specified Markov transition model is correctly specified. Further, the efficiency property is lost if the model specification is incorrect, but versions of Propositions 1 and 2 may hold if Assumption B1 is untrue. In this case, the specified Markov transition model may be viewed as "pseudo-true", and its properties may be developed from the known results on misspecified models (e.g., [24]). For example, although $\lambda_0$ is not the vector of "true" model parameters in this case, these model parameters represent the limiting MPD model based on the constraints. Thus, even if the model is not correctly specified, the MPD parameter estimator may yet converge in probability to $\lambda_0$ and exhibit asymptotically normal character. The key implication of this outcome is that the sampling properties may be used to test some relevant hypotheses that do not explicitly require the model specification to be correct. For example, we can still use an incorrectly specified Markov model that does not satisfy Assumption B1 to test the time-invariance hypothesis, $H_0 : \lambda_n = 0$, where $\lambda_n$ is the subset vector of Lagrange multipliers associated with the non-constant instruments in $\mathbf{z}_t$.

## 6. Applications of MPD Estimators

Markov models-processes recognize patterns and this opens up a range of applications in big-data economic settings such as time series observations, investment decisions under uncertainty, and life-cycle consumption and savings. For example, recent papers such as Gospodinov and Lkhagvasuren (2011) [25] and Tanaka and Toda (2013) [26] consider approximations of continuous processes with discrete Markov chains. Also, several numerical applications of MDP estimators of Markov chain models are provided by Golan, *et al.* [27]. Finally, we note that Claude Shannon, one of the developers of information theory, used Markov models to construct probabilistic models of words sent in noisy messages. Given the twenty-seven possible states (or letters, including a space), he used the Markov transition probabilities to indicate which letters are likely to follow a given letter (state) in a sequence of letters that form the words in a message. Thus, there are a wide range of possible applications of MDP estimators for finite Markov chains that range across a very broad selection of fields.

## 7. Conclusions

In this paper we have developed an easy-to-implement information theoretic basis for nonparametric estimation of dynamic econometric models from aggregate data. Estimating equations involving exogenous and endogenous variables are used as a link to the data and as a basis for modeling the Markov process. To address the unknown data sample process, we use a convex combination of members of the flexible Cressie-Read family of divergence measures to select and estimator that minimizes quadratic loss. Sampling properties of the minimum power divergence estimator are demonstrated. Application of the model is straightforward and does not involve the usual tuning parameters and regularization methods.

## Acknowledgments

We want to acknowledge the helpful comments of W. Tam Cho, E. Giacomi, A. Golan, T. Squartini and S. Villas.

## Author Contributions

The paper is joint work that reflects equal contributions of both authors.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Wissner-Gross, A.; Freer, C. Causal entropic forces. *Phys. Rev. Lett.* **2013**, *110*, doi:10.1103/PhysRevLett.110.168702.
2. Miller, D. *Behavioral Foundations for Conditional Markov Models for Aggregate Data*; Working Paper; University of Missouri: Columbia, MO, USA, 2007.
3. Gorban, A.N.; Gorban, P.A.; Judge, G. Entropy: The Markov ordering approach. *Entropy* **2010**, *12*, 1145–1193.
4. Cressie, N.; Read, T. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. Ser. B* **1984**, *46*, 440–464.
5. Rust, J. Maximum likelihood estimation of discrete control processes. *SIAM J. Control Optim.* **1988**, *26*, 1006–1024.
6. Rust, J. Structural estimation of Markov decision processes. In *Handbook of Econometrics*; Engle, R., McFadden, D., Eds.; Elsevier: Amsterdam, The Netherlands, 1994.
7. Hotz, V.J.; Miller, R.A. Conditional choice probabilities and the estimation of dynamic models. *Rev. Econ. Stud.* **1993**, *60*, 397–429.
8. Lee, T.; Judge, G.; Zellner, A. *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*, 2nd ed.; North-Holland: Amsterdam, The Netherlands, 1977.
9. MacRae, E.C. Estimation of time-varying Markov processes with aggregate data. *Econometrica* **1977**, *45*, 183–198.

10. Theil, H. A multinomial extension of the linear logit model. *Int. Econ. Rev.* **1969**, *10*, 251–259.
11. Commenges, D.; D'egout, A. A general dynamical statistical model with causal interpretation. *J. R. Stat. Soc. B* **2009**, *71*, 1–18.
12. Read, T.; Cressie, N. *Goodness-of-Fit Statistics for Discrete Multivariate Data*; Springer-Verlag: New York, NY, USA, 1988.
13. Baggerly, K. Empirical likelihood as a goodness-of-fit measure. *Biometrika* **1998**, *85*, 535–547.
14. Judge, G.; Mittelhammer, R. *An Information Approach to Econometrics*; Cambridge University Press: Cambridge, UK, 2011.
15. Judge, G.; Mittelhammer, R. Implications of the Cressie-Read family of additive divergences for information recovery. *Entropy* **2012**, *14*, 2427–2438.
16. Owen, A. *Empirical Likelihood*; Chapman and Hall: Boca Raton, FL, USA, 2001.
17. Kullback, S. *Information Theory and Statistics*; John Wiley and Sons: New York, NY, USA, 1959.
18. Gokhale, D.; Kullback. S. *The Information in Contingency Tables*; Marcel Dekker: New York, NY, USA, 1978.
19. Gorban, A.N. *Equilibrium Encircling Equations of Chemical Kinetics and Their Thermodynamic Analysis*; Nauka: Novovosibirsk, Russia, 1984.
20. Gorban, A.N.; Karlin, I.V. Family of additive entropy functions out of thermodynamic limit. *Phys. Rev. E* **2003**, *67*, 016104, doi:10.1103/PhysRevE.67.016104.
21. Newey, W.; McFadden, D. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*; Engle, R., McFadden, D., Eds.; Elsevier: Amsterdam, The Netherlands, 1994.
22. Kelton, W.; Kelton, C. Hypothesis tests for Markov process models from aggregate frequency data. *J. Am. Stat. Assoc.* **1984**, *79*, 922–928.
23. Newey, W.; West, K. A simple positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **1987**, *55*, 703–708.
24. White, H. *Asymptotic Theory for Econometricians*; Orlando: Academic Press: San Diego, CA, USA, 1984; p. 228.
25. Gospodinov, N.; Lkhagvasuren, D. A New Method for Approximating Vector Autoregressive Processes by Finite-State Markov Chains. Available online: http://mpra.ub.uni-muenchen.de /33827/1/MPRA_paper_33827.pdf (accessed on 3 November 2014).
26. Tanaka, H.; Toda, A.A. Discrete approximations of continuous distributions by maximum entropy. *Econ. Lett.* **2013**, *118*, 445–450.
27. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; Wiley: Chichester, UK, 1996.