The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

A time-series model for underdispersed or overdispersed counts

Iain L. MacDonald & Feroz Bhamani

To cite this article: Iain L. MacDonald & Feroz Bhamani (2018): A time-series model for underdispersed or overdispersed counts, The American Statistician, DOI: 10.1080/00031305.2018.1505656

To link to this article: https://doi.org/10.1080/00031305.2018.1505656

A time-series model for underdispersed or overdispersed counts

Iain L. MacDonald
Centre for Actuarial Research
University of Cape Town
and
Feroz Bhamani
African Institute of Financial Markets and Risk Management
University of Cape Town

July 7, 2018

Abstract

It is common for time series of unbounded counts (that is, nonnegative integers) to display overdispersion relative to the Poisson. Such an overdispersed series can be modeled by a hidden Markov model with Poisson state-dependent distributions (a 'Poisson–HMM'), since a Poisson–HMM allows for both overdispersion and serial dependence. Time series of underdispersed counts seem less common, but more awkward to model; a Poisson–HMM cannot cope with underdispersion. But if in a Poisson–HMM one replaces the Poisson distributions by Conway–Maxwell–Poisson distributions, one gets a class of models which can allow for under- or overdispersion (and serial dependence). In addition, this class can cope with the combination of slight overdispersion and substantial serial dependence, a combination that is apparently difficult for a Poisson–HMM to represent.

We discuss the properties of this class of models, and use direct numerical maximization of likelihood to fit a range of models to three published series of counts which display underdispersion, and to a series which displays slight overdispersion plus substantial serial dependence. In addition, we illustrate how such models can be fitted without imputation when some observations are missing from the series, and how approximate standard errors of the parameter estimates can be found.

Keywords: Conway–Maxwell–Poisson distribution; hidden Markov model; missing data

1 Introduction

Many time series of counts display overdispersion relative to the Poisson, that is, the sample variance exceeding the sample mean. This phenomenon has received considerable attention in the literature; see e.g. Yang *et al.* (2015), who write that such series 'may exhibit three distinctive features: overdispersion, zero-inflation and temporal correlation.' Safari *et al.* (2017) make several references to overdispersion in series of counts, but do not mention the possibility of underdispersion. But there are cases of underdispersion, e.g. the series of counts of pedestrians studied by Fürth (1918). A (stationary) Poisson–hidden Markov model $\{X_t\}$ can allow for both overdispersion and serial dependence, but is unable to allow for underdispersion; we shall see in Section 2.2 that, for such a model,

$$\mathrm{Var}(X_t) = \mathrm{E}(X_t) + \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2.$$

Here $\boldsymbol{\delta}$ is the stationary distribution of the underlying Markov chain, and μ_i is the ith state-dependent mean, i.e. the mean (and variance) of X_t conditional on the Markov chain being in state i.

If, however, one replaces the Poisson as state-dependent distribution by a distribution which is or can be underdispersed relative to the Poisson, one can achieve underdispersion in X_t, i.e. $\mathrm{Var}(X_t) < \mathrm{E}(X_t)$. The Conway–Maxwell–Poisson (CMP) is one such distribution, and a stationary hidden Markov model (HMM) with CMP state-dependent distributions can accommodate both underdispersion and serial dependence. Furthermore, it can if necessary accommodate equi- or overdispersion, although those are not the focus here.

We describe this class of models, and use maximum likelihood to fit such models to three published series of counts which display underdispersion, and also to a series which displays slight overdispersion plus substantial serial dependence; this is a combination which a Poisson–HMM will probably struggle to represent. We describe two of the analyses in some detail, that relating to the pedestrian counts of Fürth, and that relating to the gold particle counts of Westgren (1916). We confine our attention to series which can reasonably be treated as stationary, although there are several ways in which HMMs can be modified in order to incorporate nonstationarity. Throughout this paper, any reference to under- or overdispersion is to be understood as relative to the Poisson.

We begin with a fairly detailed introduction to HMMs, their properties and their estimation by maximum likelihood. After a brief discussion of the Conway–Maxwell–Poisson distribution, we describe its use as a state-dependent distribution in an HMM. The computational methods are then described, and the four data analyses and a concluding discussion follow.

2 Hidden Markov models

2.1 Structure

Hidden Markov models are well known for their applications in speech processing and bioinformatics, but may be unfamiliar to readers as general-purpose models for time series. We therefore provide here an introduction to such models. A fuller account is provided by Zucchini *et al.* (2016).

In a (discrete-time) hidden Markov model $\{X_t\}$ it is assumed that there is a latent irreducible Markov chain $\{C_t\}$, the current state of which determines the distribution of the observation. In many applications the Markov chain is on a very small number (m) of states, e.g. two or three. It is often assumed that the Markov chain is stationary, and that, conditional on the Markov chain, the observations are independent. We make these two assumptions here, but they can be relaxed.

To complete the specification of such a model we need, in addition to the Markov chain, a distribution for the observations in each of the states; these are the m 'state-dependent distributions', and can be any distributions appropriate to the nature of the observations, e.g. Poisson distributions, negative binomial distributions, or (in another context) normal distributions. Such distributions need not all come from the same family. For instance, in a two-state HMM the distribution could be Poisson in state 1 and identically zero in state 2. The resulting model is a serially dependent process with a zero-inflated Poisson as the marginal distribution.

An HMM, in its most basic form, is conveniently depicted by the directed graph shown in Figure 1. The horizontal arrows display the Markov-dependence of the latent variables C_t. The special case in which those arrows are absent represents an independent mixture

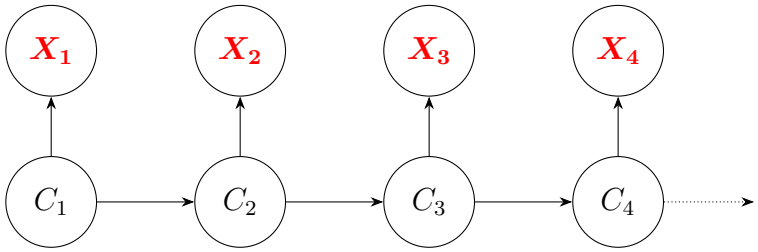

Figure 1: Directed graph of basic HMM. The random variables X_t represent the observable quantities.

of distributions, as opposed to a 'Markov-dependent mixture'; the latter is an alternative name for an HMM.

2.2 Moments

We now state several useful results pertaining to the moments of such a model, results which appear (for instance) in the exercises of Zucchini *et al.* (2016, pp. 23, 42) or are minor extensions of results appearing there. Proofs are outlined in the appendix.

For general state-dependent distribution, and for $m \in \mathbb{N}$ states,

$$\mathrm{Var}(X_t) = \sum_{i=1}^{m} \delta_i \sigma_i^2 + \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2. \tag{1}$$

(In passing, we note that this equation holds for an independent mixture of m distributions, since such a mixture is a special case of an HMM.) If all the state-dependent distributions are Poisson, Equation (1) reduces to

$$\mathrm{Var}(X_t) = \sum_{i=1}^{m} \delta_i \mu_i + \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2 = \mathrm{E}(X_t) + \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2. \tag{2}$$

In addition we have in general that, for $k \in \mathbb{N}$,

$$\mathrm{Cov}(X_t, X_{t+k}) = \boldsymbol{\delta} \boldsymbol{M} \boldsymbol{\Gamma}^k \boldsymbol{\mu}' - (\boldsymbol{\delta} \boldsymbol{\mu}')^2. \tag{3}$$

Here and elsewhere we use the following notation:

- the mean and variance of the ith state-dependent distribution are μ_i and σ_i^2,

- m is the number of states in the Markov chain,
- $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$,
- $\boldsymbol{M}$ is diag($\boldsymbol{\mu}$), the diagonal matrix having $\boldsymbol{\mu}$ as principal diagonal,
- $\mathbb{N}$ denotes the positive integers,
- the prime symbol denotes transposition, and
- the row vector $\boldsymbol{\delta}$ is the (unique, strictly positive) stationary distribution implied by the $m \times m$ transition probability matrix (TPM) $\boldsymbol{\Gamma} = (\gamma_{ij})$. We use the common convention that the *row* sums of the TPM are 1.

The autocorrelation function (ACF) then follows, for $k \in \mathbb{N}$, as

$$\rho(k) = \text{Cov}(X_t, X_{t+k})/\text{Var}(X_t). \tag{4}$$

For the case $m = 2$, this reduces to

$$\rho(k) = Aw^k, \tag{5}$$

where $w = 1 - \gamma_{12} - \gamma_{21}$ and

$$A = \frac{\delta_1\delta_2(\mu_1 - \mu_2)^2}{\delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2}. \tag{6}$$

Notice that, apart from degenerate cases, $A \in (0, 1)$. Notice also that the ACF in (5) is more general in form than that of an autoregressive process of order 1, whether a Gaussian AR(1) or an INAR(1), the integer-valued AR(1) used by Weiß (2013), for instance. In the latter case, negative autocorrelation is not possible, but here it is, since the range of w is $(-1, 1)$.

Although Equations (1), (3) and (4) are in general sufficient for the numerical computation of the ACF of an HMM, it is possible to find an expression for $\rho(k)$ in terms of powers of the eigenvalues (other than 1) of the TPM $\boldsymbol{\Gamma}$. In fact, Equation (5) is precisely that expression for the case $m = 2$. More generally, we proceed as follows: assume that $\boldsymbol{\Gamma}$ is diagonalizable — a reasonable assumption — and write it as

$$\boldsymbol{\Gamma} = \boldsymbol{U}\text{diag}(1, \omega_2, \ldots, \omega_m)\,\boldsymbol{U}^{-1}. \tag{7}$$

(It is convenient to scale the first column of $\boldsymbol{U}$ to be a column of ones.) Equations (3) and (7) imply that

$$\mathrm{Cov}(X_t, X_{t+k}) = \boldsymbol{\delta M U}\,\mathrm{diag}(1, \omega_2^k, \ldots, \omega_m^k)\,\boldsymbol{U}^{-1}\boldsymbol{\mu}' - (\boldsymbol{\delta\mu}')^2. \tag{8}$$

Now define $\boldsymbol{a} = \boldsymbol{\delta M U}$ and $\boldsymbol{b}' = \boldsymbol{U}^{-1}\boldsymbol{\mu}'$, and it follows that, for $k \in \mathbb{N}$,

$$\mathrm{Cov}(X_t, X_{t+k}) = a_1 b_1 + \sum_{i=2}^{m} a_i b_i \omega_i^k - (\boldsymbol{\delta\mu}')^2 = \sum_{i=2}^{m} a_i b_i \omega_i^k. \tag{9}$$

We use this expression in the examples of Section 2.4 and elsewhere.

2.3 Maximum-likelihood estimation and standard errors

In order to fit an HMM by maximum likelihood, one can make use of a convenient expression giving the likelihood as a product of matrices, and maximize it numerically with respect to all the parameters. That expression appears as Equation (2.12) of Zucchini *et al.* (2016, p. 37) and in many other places, but it may be useful to the reader if we repeat it here. Let L_T denote the likelihood of T consecutive observations x_1, x_2, …, x_T. Let p_i denote the probability mass or density function of the ith state-dependent distribution (of m), and let $\boldsymbol{P}(x)$ denote the diagonal matrix with principal diagonal $(p_1(x), \ldots, p_m(x))$. Then

$$L_T = \boldsymbol{\delta P}(x_1)\boldsymbol{\Gamma P}(x_2)\boldsymbol{\Gamma P}(x_3)\cdots\boldsymbol{\Gamma P}(x_T)\mathbf{1}'. \tag{10}$$

Here $\boldsymbol{\Gamma}$ and $\boldsymbol{\delta}$ are as before, and $\mathbf{1}'$ is a column vector of ones.

With the expression (10) for L_T available, the log-likelihood of T observations can be evaluated in $O(Tm^2)$ operations, which is not *a priori* obvious and is a great improvement on the $O(Tm^T)$ operations apparently needed if one considers only a certain multiple-sum expression for the likelihood. (Indeed, there are statements in the literature of the 1990s that likelihood evaluation of a basic HMM is computationally infeasible, even for two states and moderately large T. Fortunately, that is not true. See for instance Albert (1991) and the 'reader reaction' of Le *et al.* (1992).)

Provided that the state-dependent probabilities can be evaluated efficiently, the log-likelihood can therefore be evaluated routinely, except that some simple precautions have to be taken against numerical underflow; see Zucchini *et al.* (2016, p. 48), who describe

how this can be achieved by a scaling of the likelihood computation. In addition, the optimization must be performed in such a way as to respect the constraints that there will be on parameters, e.g. the row-sum constraints on the TPM. One way of implementing the constraints is to transform the constrained 'natural parameters' to unconstrained 'working parameters', and then to use a standard unconstrained optimizer to maximize the log-likelihood with respect to the working parameters. That is the approach followed here, with the modified Newton optimizer `nlm` provided by R (R Core Team, 2017) being the one used. However, it needs to be noted that multiple local optima of the likelihood are possible; there is no guarantee that a local optimum that has been found is the global optimum. It is wise to try multiple sets of starting values for the optimizations, especially if there are many parameters.

In such models, it is possible to arrive at approximate standard errors of the maximum likelihood estimators (MLEs) of the natural parameters by making use of the numerical Hessian available from `nlm`, and this is done in several of the examples: see Sections 5.5 and 8.3. If one has optimized over working parameters rather than natural, however, the Hessian available will not be the appropriate one. It will be the Hessian with respect to the working parameters, and we need the Hessian with respect to the natural parameters if we seek standard errors for the natural parameters. This can be remedied by re-running the optimization without constraints, with starting values at or close to the optimum already found; that is the technique used in this work, although it is not the only possibility. Such Hessians, being based on two numerical differentiations, cannot be expected to be very accurate. Furthermore, the standard errors will (as always) not be useful in constructing confidence intervals if parameter estimates are on the boundary of the parameter space or it is unrealistic to assume normality of the MLEs.

2.4 Examples: Poisson–HMMs with 2–4 states

As simple illustrative examples we present here Poisson–HMMs with 2–4 states fitted (as described in Section 2.3) to a series of length 242 representing weekly sales (in integer units) of a soap product. The data appear in full in Zucchini *et al.* (2016, p. 25), and the source is `http://research.chicagobooth.edu/kilts/marketing-databases/dominicks`, a data-

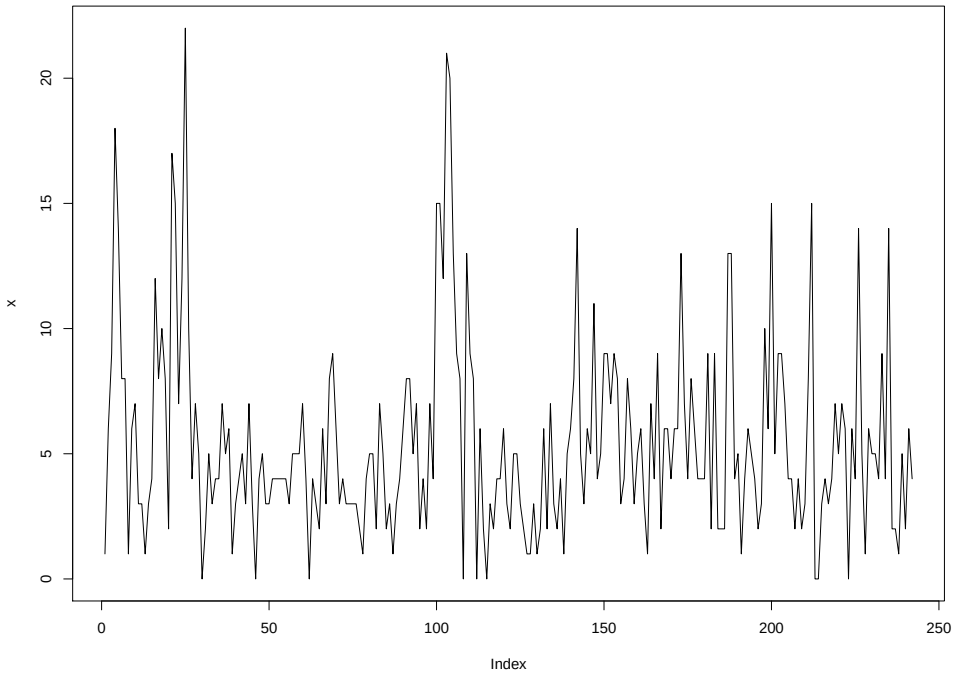

Figure 2: Weekly sales of a soap product.

base of the Kilts Center for Marketing, University of Chicago. (The product was 'Zest White Water 15 oz.', with code 3700031165, and the store number 67.) For a plot, see Figure 2. The series displays both overdispersion and serial dependence: the sample mean, variance and first-order autocorrelation are 5.44, 15.40 and 0.39 respectively.

The two-state model has means $\boldsymbol{\mu} = (4.02, 11.37)$ and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.912 & 0.088 \\ 0.370 & 0.630 \end{pmatrix},$$

which implies the stationary distribution $\boldsymbol{\delta} = (0.809, 0.191)$. The model mean and variance are 5.43 and 13.78. The ACF of the model is, for $k \in \mathbb{N}$, $\rho(k) = 0.606 \times 0.542^k$.

The comparable three-state model — which also appears in MacDonald and Zucchini (2016) — has means $\boldsymbol{\mu} = (3.74, 8.44, 14.93)$, stationary distribution $\boldsymbol{\delta} = (0.722, 0.220, 0.058)$

and

$$\mathbf{\Gamma} = \begin{pmatrix} 0.864 & 0.117 & 0.019 \\ 0.445 & 0.538 & 0.017 \\ 0.000 & 0.298 & 0.702 \end{pmatrix}.$$

The model mean and variance are 5.42 and 14.72. The model ACF is, for $k \in \mathbb{N}$,

$$\rho(k) = 0.539 \times 0.682^k + 0.0926 \times 0.422^k.$$

Finally, the comparable four-state model has means $\boldsymbol{\mu} = (0.00, 3.87, 8.20, 14.69)$, stationary distribution $\boldsymbol{\delta} = (0.021, 0.701, 0.214, 0.065)$ and

$$\mathbf{\Gamma} = \begin{pmatrix} 0.177 & 0.634 & 0.000 & 0.189 \\ 0.000 & 0.881 & 0.099 & 0.020 \\ 0.079 & 0.329 & 0.567 & 0.025 \\ 0.000 & 0.000 & 0.360 & 0.640 \end{pmatrix}.$$

The model mean and variance are 5.42 and 14.90. The model ACF is, for $k \in \mathbb{N}$,

$$\rho(k) = 0.539 \times 0.660^k - 0.027 \times 0.338^k + 0.124 \times 0.266^k.$$

Examination of the stationary distributions and the vectors of state-dependent means suggests that the four-state model is roughly the three-state model with state 1 of the latter split into two states, one of which allows only the observation zero.

The reader may reasonably question whether the extra complexity of the four-state model (which has 16 parameters) is worthwhile. One way of answering such model-selection questions is to use Akaike's information criterion (AIC) or the Bayesian information criterion (BIC), both of which are discussed by Zucchini (2000). These criteria trade off any improvement in (log-) likelihood against the number of parameters needed to achieve such improvement. Using (*inter alia*) AIC and BIC, MacDonald and Zucchini (2016) concluded that the four-state model was inferior to the two- and three-state models, and that there was little to choose between the latter two. Furthermore, a comparable one-state model (i.e. a sequence of independent Poisson random variables) showed up as much inferior to the other three.

3 Hidden Markov models using the CMP distribution

3.1 The Conway–Maxwell–Poisson distribution

The Conway–Maxwell–Poisson distribution dates back (at least) to the work of Conway and Maxwell (1961), who showed that it arises very naturally in a queueing context. In an important paper published decades later, Shmueli *et al.* (2005) established many of its properties and demonstrated its general usefulness as a model for unbounded counts (nonnegative integers). In such a distribution the probability mass function is of the form

$$\Pr(X = x) \propto \lambda^x/(x!)^\nu$$

for all nonnegative integers x. More fully,

$$\Pr(X = x) = \frac{1}{Z(\lambda, \nu)} \frac{\lambda^x}{(x!)^\nu},$$

where $Z(\lambda, \nu) = \sum_{k=0}^{\infty} \lambda^k/(k!)^\nu$. The parameters are $\lambda > 0$ and $\nu \geq 0$, but if $\lambda \geq 1$ and $\nu = 0$ the distribution is undefined. As Shmueli *et al.* point out, two special cases and a limit of the CMP are familiar distributions. The case $\nu = 1$ yields the Poisson distribution with mean λ, the case $\nu = 0$ and $0 < \lambda < 1$ the geometric with 'success probability' $1 - \lambda$, and the limit as ν tends to ∞ the Bernoulli with success probability $\lambda/(\lambda + 1)$. Underdispersion is indicated by $\nu > 1$, and overdispersion by $\nu < 1$. A simple result linking the parameters, due to R. Snyder, is that, for a random variable X having a CMP distribution, $\mathrm{E}(X^\nu) = \lambda$ (Sellers and Shmueli, 2010, p. 946).

Because of the infinite sum in the normalizing constant of a CMP distribution, it is not entirely straightforward to evaluate the probability mass function, but in this work we need to do so repeatedly. The R package `COMPoissonReg` (Sellers *et al.*, 2017) was used for this purpose, in particular the function `dcmp`.

Several innovative uses of CMP distributions have appeared recently. Some of these are as follows. Zhu (2012b) has used integer-valued GARCH models with CMP conditional distribution to model time series of counts. Wu *et al.* (2013) discuss Bayesian models for spatio-temporal count data based on CMP distributions. Sur *et al.* (2015) have fitted independent mixtures of CMP distributions to several data-sets. The models we shall describe in Section 3 are Markov-dependent mixtures of CMP distributions, and thereby

more general. Zhu *et al.* (2017) have proposed the use of the CMP distribution to generalize the homogeneous Poisson process on the real line. Huang (2017) has explored regression models in which the response has a CMP distribution parametrized by its mean and ν, rather than by λ and ν, and has provided three applications thereof.

3.2 The Conway–Maxwell–Poisson as state-dependent distribution

One of the many ways in which HMMs can be modified or 'customized' is the use of state-dependent distributions that are, or seem to be, particularly appropriate to the data for which a model is sought. It was suggested by MacDonald (2017) that CMP distributions could be used as state-dependent distributions in an HMM, and that suggestion is pursued here. What that involves (compared to a Poisson–HMM) is the replacement of the Poisson state-dependent distributions by CMP distributions. The resulting model consists of the TPM $\mathbf{\Gamma}$ and, for each of the m states, the two parameters λ_i and ν_i; there are in all $(m^2 - m) + 2m = m(m+1)$ parameters to be determined. In what follows, we use the row vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ to denote the state-dependent parameters of such a model.

We saw at (1) that, for any HMM $\{X_t\}$,

$$\mathrm{Var}(X_t) = \sum_{i=1}^{m} \delta_i \sigma_i^2 + \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2.$$

If, for all i, $\sigma_i^2 < \mu_i$, then $\sum_i \delta_i \sigma_i^2 < \sum_i \delta_i \mu_i = \mathrm{E}(X_t)$, and it is possible that $\mathrm{Var}(X_t) < \mathrm{E}(X_t)$ or $\mathrm{Var}(X_t) \geq \mathrm{E}(X_t)$; both cases can occur. Hence an HMM with CMP state-dependent distributions (a 'CMP–HMM') is a candidate as a model for time series of counts, whether underdispersed or overdispersed. Notice that a mixture — independent or Markov-dependent — of underdispersed distributions may itself be overdispersed; this is likely if at least some of the component means are widely spaced and the sum $\sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2$ is therefore large. Furthermore, a mixture of two CMP distributions with one ν below 1 and the other above 1 may be either under- or overdispersed. For an extensive discussion of the effect of mixing on dispersion levels, see Sellers and Shmueli (2013).

The fitting of a CMP–HMM by direct numerical maximization of likelihood can proceed essentially as in the case of a Poisson–HMM; see Section 2.3 of this paper or Zucchini

et al. (2016, Ch. 3). That is, a function is written to evaluate $-\log L_T$ as a function of unconstrained working parameters by means of (10), and then that function is minimized by means of `nlm` or some other unconstrained optimizer.

4 Computational methods

All the HMMs that will be reported in Sections 5–8 were fitted by direct numerical maximization of log-likelihood, as described in Section 2.3. The R code of Zucchini *et al.* (2016, Appendix A) was used or extended for this purpose. That code is also available at `http://www.hmms-for-time-series.de/second/appendix-a/ZMLcode.txt`. Our code to fit CMP–HMMs and other, simpler, models accompanies this paper.

The main extension needed is that the code has to provide for CMP state-dependent distributions in addition to Poisson, and the routine `dcmp` (in the package `COMPoissonReg`) was used here with default settings. A log transformation was used to impose positivity on the parameters λ_i and ν_i, in the same way as a log transformation can be used to constrain the state-dependent means in a Poisson–HMM. Another extension is that optimizations (with respect to natural parameters) are re-run without constraints in order to find the Hessian with respect to the natural parameters and so find approximate standard errors.

Once a model has been fitted, it is of course informative to find the model mean and variance, and that requires deducing the mean and variance of a CMP distribution from the estimates of λ and ν. There are approximate formulas for the mean and variance: $\lambda^{1/\nu} - \frac{\nu-1}{2\nu}$ and $\nu^{-1}\lambda^{1/\nu}$ respectively. We did not use these approximations, which are reported to be accurate if $\nu < 1$ or $10^\nu < \lambda$ (Sellers *et al.*, 2012). Apart from the routines `com.mean` and `com.var` provided by the package `compoisson` (Dunn, 2012), there are several possible approaches. One is just taking a very large number of terms in $\frac{1}{Z(\lambda,\nu)} \sum_x x\lambda^x/(x!)^\nu$, and similarly for the variance. That is the approach we followed. Another, perhaps less obvious, approach is to solve the following equations (numerically) for μ and then σ^2:

$$0 = g(\mu) = \sum_{x=0}^{\infty} (x - \mu) \frac{\lambda^x}{(x!)^\nu} \tag{11}$$

and

$$0 = h(\sigma^2) = \sum_{x=0}^{\infty} \left((x-\mu)^2 - \sigma^2\right) \frac{\lambda^x}{(x!)^\nu}. \tag{12}$$

Here again it is necessary to truncate the series involved to some large number of terms, but, even with thousands of terms, evaluation of the sums in Equations (11) and (12) can be performed fast enough to make iterative solution (by the R routine `uniroot`) possible. This second approach was used to check a sample of the values produced by the first.

5 The pedestrian counts data of Fürth (1918), and the models fitted

5.1 The data

Fürth (1918, Tabelle I) presented a series of 505 counts of the number of pedestrians traversing a city block, the counts being taken at 5-second intervals.[1] According to Fürth (1919) the block was 24 m wide, and the estimated speed of pedestrians 5 km h^{-1}, which implies about 17 s for the 24 m. The sample mean count is 804/505=1.592, and the sample variance (with denominator 504) 1.508; so there is slight underdispersion, which makes it unlikely that a necessarily overdispersed model such as a Poisson–HMM will be adequate. Nevertheless, a Poisson–HMM will be one of the models fitted. There seems to be substantial autocorrelation at low lags: see Figure 3. The first two autocorrelations are 0.665 and 0.322.

This series has also been discussed by (*inter alios*) Jung and Tremayne (2006b) and Martin *et al.* (2014); the latter authors state of these data that prior to their paper 'no fully satisfactory time-series model seems to have been unearthed', and therefore fit a wide range of models of INARMA type. These are integer-valued time series models broadly analogous to Gaussian autoregressive moving-average models, but based on the idea of binomial thinning. Because these models incorporate a moving-average component, maximum-likelihood methods are apparently not feasible, and a moment-based estimation technique is used.

[1] We use exactly those 505 counts. Heyde and Seneta (1972) add a 506th count, equal to zero; they write that 'observation stopped with 506th reading zero.'

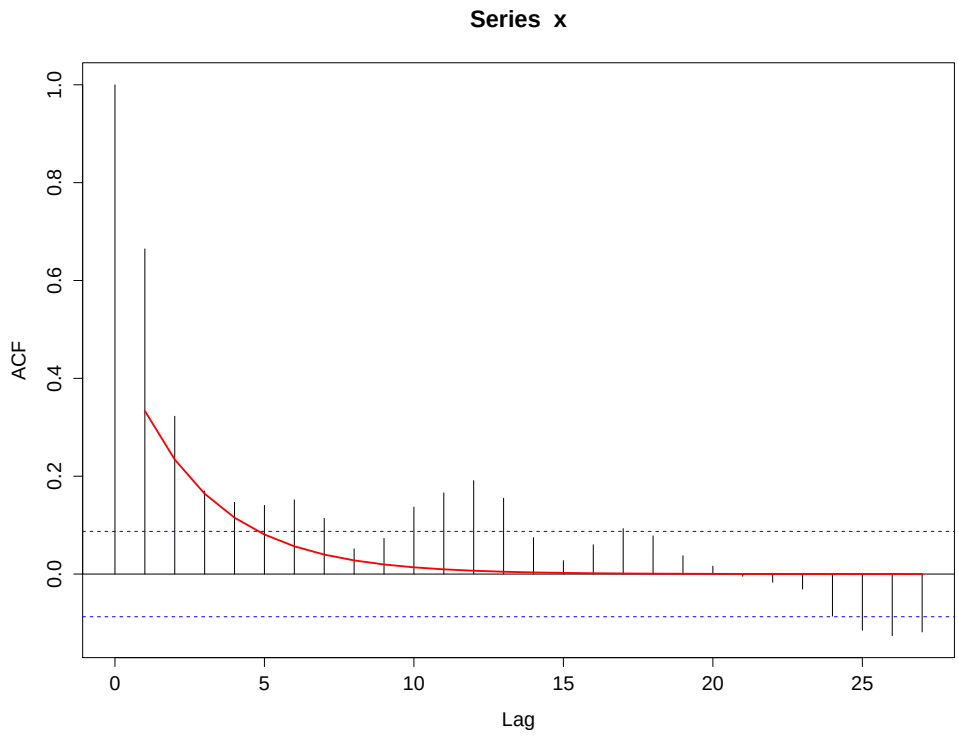

Figure 3: ACF of Fürth's data, with that of the modified two-state CMP–HMM shown as a continuous (red) line.

Here we discuss rather different models. We present three models of HMM type, fitted by maximum likelihood, but include for comparison three other models which can be expected to be unsatisfactory, because they do not allow for serial dependence. The most complex model considered is a two-state stationary HMM with a CMP distribution associated with each state, a model which has six parameters. It is a modified but essentially equivalent version of that model which, of those considered here, emerges as 'best' by the criteria AIC and BIC.

5.2 Three models assuming independence of observations

Independent observations on a single Poisson

A Poisson distribution with mean 1.592 is the simplest model one could attempt, and yields log-likelihood -785.8321. It captures neither underdispersion nor serial dependence.

Independent observations on a single CMP

A single CMP distribution was fitted by numerical maximization of the (log)-likelihood. The estimated λ and ν are 1.715 and 1.091; by the latter the model is, as one would expect, slightly underdispersed. But the serial dependence is not captured.

Independent mixture of two Poissons

The fitted model degenerates to a mixture of two Poissons with the same mean, 1.592, thereby eliminating overdispersion from the model. As the data are underdispersed, this is not surprising.

5.3 A two-state Poisson–HMM

The model fitted is $\boldsymbol{\mu} = (0.2545, 1.9014)$ and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8313 & 0.1687 \\ 0.0401 & 0.9599 \end{pmatrix},$$

which implies the stationary distribution $\boldsymbol{\delta} = (0.1919, 0.8081)$. The model mean and variance are 1.585 and 2.006, so clearly the underdispersion is not captured.

5.4 A two-state HMM with CMP state-dependent distributions

The model fitted is $\boldsymbol{\lambda} = (0.8862, 9.165)$, $\boldsymbol{\nu} = (28.75, 2.400)$, and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8086 & 0.1914 \\ 0.1070 & 0.8930 \end{pmatrix}.$$

The corresponding stationary distribution is $\boldsymbol{\delta} = (0.3586, 0.6414)$. The mean and variance of this model are 1.585 and 1.463, so there is indeed underdispersion in the model. As the associate editor has pointed out, however, the distribution in state 1 has an extremely high value of ν. It is very close indeed to a Bernoulli distribution. Consequently, the model can be described as essentially an HMM with a Bernoulli distribution in one state and an underdispersed CMP distribution in the other. We therefore fitted a modified model of precisely that kind, which we now describe.

5.5 Modified two-state CMP–HMM

The modified model consists of a stationary two-state Markov chain, with a Bernoulli distribution in state 1 and a CMP distribution in state 2. The model of Section 5.4 provides excellent starting values for the parameters, but other starting values were also attempted, and all yielded the same fitted model, with maximized log-likelihood -715.5128. That fitted model has $\boldsymbol{\Gamma}$ (and consequently $\boldsymbol{\delta}$) as in Section 5.4, but we state all the parameter estimates here, with associated approximate standard errors in brackets. The TPM is

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8086 & 0.1914\,(0.0332) \\ 0.1070\,(0.0268) & 0.8930 \end{pmatrix}.$$

In state 1 the observation has a Bernoulli distribution with probability $0.4698\,(0.0592)$ of taking the value 1. In state 2 the observation has a CMP distribution with parameters $\lambda_2 = 9.165\,(2.815)$ and $\nu_2 = 2.400\,(0.256)$.

The mean and variance of the model are again 1.585 and 1.463; the sample mean and variance are 1.592 and 1.508. The frequencies expected under the model, and the actual frequencies observed, are as shown in Table 1. From that table we note that, although the model reproduces the sample mean and variance fairly closely, the frequencies do not

Table 1: Comparison of actual and expected frequencies, modified two-state CMP–HMM for pedestrian counts of Fürth.

	0	1	2	3	4	5	6	7	8
Expected	104.0	158.0	126.6	83.1	27.3	5.3	0.7	0.1	0.0
Actual	98	165	136	70	26	8	1	1	0

Table 2: Comparison of models for pedestrian counts of Fürth.

Model	$-l$	k	AIC	BIC
(Indep. observations on) a single Poisson	785.8321	1	1573.7	1578
Single CMP	785.4309	2	1574.9	1583
Independent mixture of two Poissons	785.8321	3	1578	1590
Two-state Poisson–HMM	747.5185	4	1503	1520
Two-state CMP–HMM	715.5128	6	1443	1468
Modified two-state CMP–HMM	715.5128	5	1441	1462

Here and in similar tables, l denotes the maximized log-likelihood and k the number of parameters estimated, AIC $= -2l + 2k$, and BIC $= -2l + k \times \log(\text{no. of observations})$.

correspond as well. The ACF of the model is, for $k \in \mathbb{N}$,

$$\rho(k) = 0.3335 \times 0.7017^{k-1} = 0.4754 \times 0.7017^{k}.$$

This does not match the sample ACF particularly well (see Figure 3), but it certainly implies substantial autocorrelation at low lags. If one considers this modified model to have five parameters rather than six, it is by AIC and BIC the best of those considered; see Table 2.

5.6 Discussion

It is tempting to devise a substantive interpretation for the states of an HMM, and very satisfying if one can find a convincing interpretation, but HMMs (and latent-variable models in general) seem prone to over-interpretation, and we prefer to resist the temptation here. A model can be useful even if merely an empirical one, as opposed to a substantive one.

By AIC and BIC the CMP–HMMs are a big improvement on the Poisson–HMM; this is also not surprising, given the overdispersion which is built into a Poisson–HMM (in other than trivial degenerate cases). The extra parameter(s) seem well justified. It is difficult to compare the CMP–HMMs with the models fitted by Martin *et al.* (2014), as likelihood values are not available for their models. However, in view of the actual and expected frequencies and the ACF, our (modified) CMP–HMM can probably also be described as not fully satisfactory. But one should not over-emphasize the importance of similarity in model and sample ACFs, as little or nothing is known about the finite-sample behavior of the sample ACF in this context.

6 The IP counts of Weiß (2007, 2008)

6.1 The data

A published series known to exhibit slight underdispersion is the 'IP counts' series of Weiß (2007, 2008), who investigated the number of different IP addresses registered within periods of two minutes' length, between 10 A.M. and 6 P.M., at a server at the University of Würzburg. The series is of length 241. The data used here, relating to 29th November 2005, were read from Zhu (2012a, Fig. 2). The sample mean, variance and first-order autocorrelation are (respectively) 1.286, 1.205 and 0.292. This series has also been discussed by Martin *et al.* (2014), who concluded that there is little to choose between an integer-valued AR(1) model and an integer-valued MA(1).

6.2 A two-state HMM with CMP state-dependent distributions

The model fitted is $\boldsymbol{\lambda} = (1.120, 29.46)$, $\boldsymbol{\nu} = (1.770, 3.363)$, and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8721 & 0.1279 \\ 0.2923 & 0.7077 \end{pmatrix}.$$

The corresponding stationary distribution is $\boldsymbol{\delta} = (0.6957, 0.3043)$. The mean and variance of this model are 1.281 and 1.192, so the model captures the underdispersion. The ACF of the model is $\rho(k) = 0.4333 \times 0.5798^k$.

7 The emergency counts series of Weiß (2013)

7.1 The data

This series appears in Weiß (2013, Fig. 5), who attributes the data to Prof. Murat Testik of Hacettepe University, Ankara. The data were collected in the examination room of the emergency department of a children's hospital, and consist of 96 observations, at 10-minute intervals, of the number of patients between call for examination and first treatment. The sample mean, variance and first-order autocorrelation are 2.562, 1.870 and 0.664. Weiß concludes that the dependence structure is of AR(1) type.

7.2 A two-state HMM with CMP state-dependent distributions

The model fitted is $\boldsymbol{\lambda} = (6.2971, 428.45)$, $\boldsymbol{\nu} = (3.076, 4.415)$, and

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.9182 & 0.0818 \\ 0.0627 & 0.9373 \end{pmatrix}.$$

The corresponding stationary distribution is $\boldsymbol{\delta} = (0.4337, 0.5663)$. The mean and variance of this model are 2.642 and 1.841, so the model certainly captures the underdispersion. The ACF of the model is $\rho(k) = 0.5831 \times 0.8555^k$; see Figure 4, which displays the relatively slow decline of the ACF of the HMM.

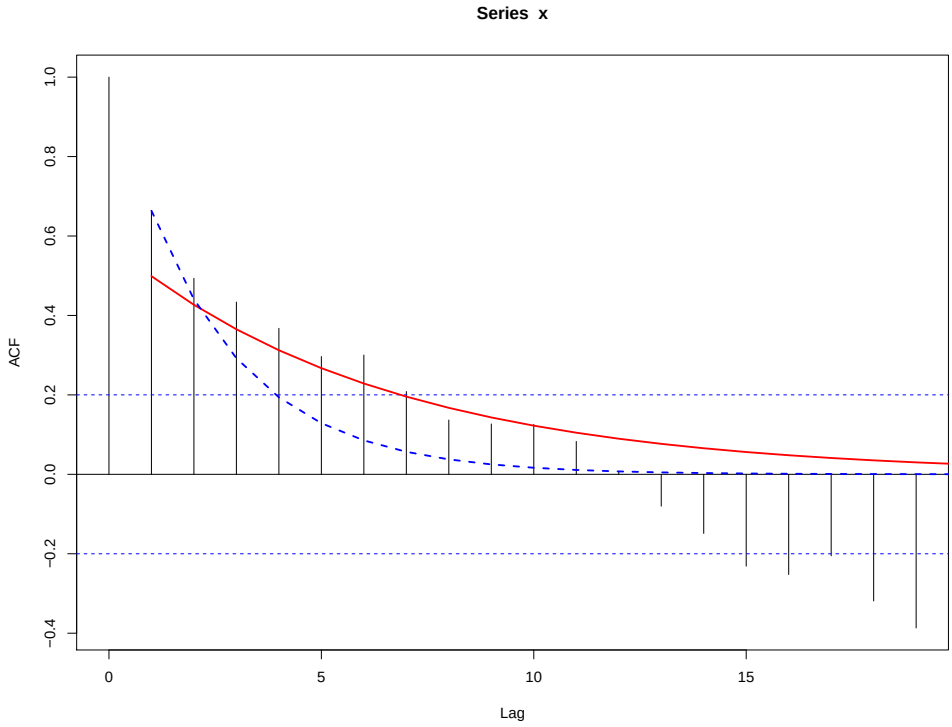

Figure 4: ACF of emergency counts, with that of the CMP–HMM as a continuous (red) line, and an AR(1) ACF (0.664^k) as a dashed (blue) line. The HMM is better able to represent slow decline in the autocorrelations.

8 The gold particle counts of Westgren (1916)

In a Poisson–HMM, the overdispersion and the serial dependence arise from the same source, the Markov-dependent mixing of the Poisson distributions. It seems unlikely that such a model can simultaneously accommodate very slight overdispersion and substantial serial dependence. (In the two-state case this is clear from the overdispersion $\mathrm{Var}(X_t) - \mathrm{E}(X_t) = \delta_1\delta_2(\mu_1-\mu_2)^2$ and Equations (5) and (6).) A series of gold particle counts originally published by Westgren (1916) seems to be a case in point, either the subseries analysed by Jung and Tremayne (2006a) or the full series as published by Guttorp (1991).

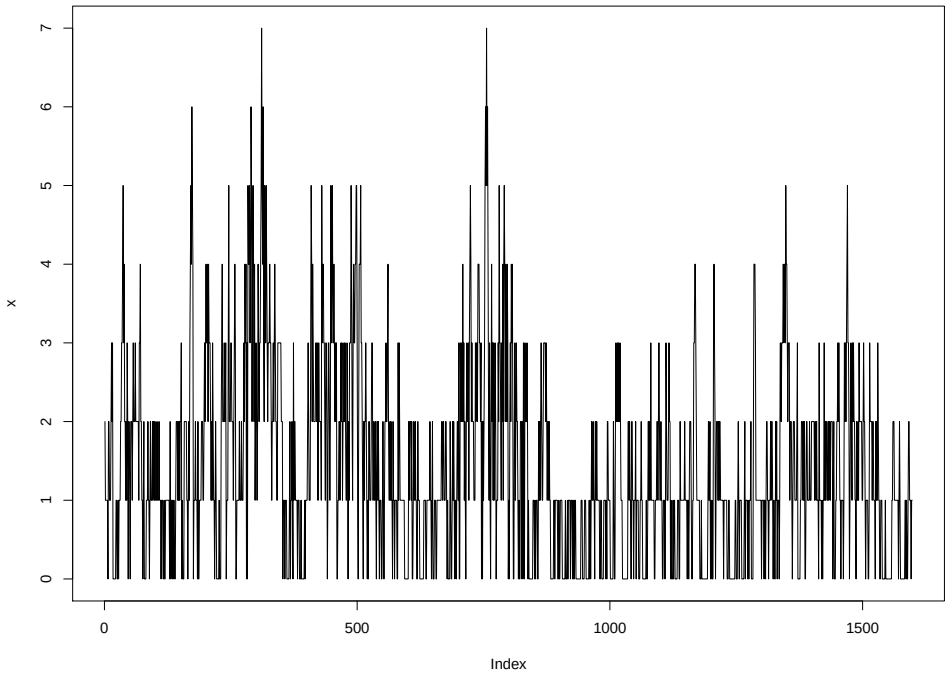

Figure 5: Guttorp's version of the gold particle counts.

8.1 The data

Guttorp (1991, pp. 191–192) gives in full a series of length 1598, pointing out which 11 observations are missing and have been imputed by him, and how, and noting that these data arise from experiment C of Westgren (1916); see Figure 5. The 380 counts of Jung and Tremayne (2006a) appear to be exactly observations 502–881 as in Guttorp, hence with four missing observations as imputed by Guttorp. The sample mean and variance of the 380 counts are 1.56 and 1.62, and the first-order sample autocorrelation is 0.57. The corresponding figures for the 1598 counts (with imputation) are 1.42, 1.51 and 0.64.

A point which is worth stressing and is relevant here is that, if one wishes (only) to fit an HMM by maximum likelihood, it is entirely unnecessary to impute missing observations. It is possible to evaluate the likelihood of just those observations actually available (Zucchini *et al.*, 2016, p. 40) and maximize that likelihood. (If the observation at time t is missing, the matrix $\boldsymbol{P}(x_t)$ is simply replaced in Equation (10) by the identity matrix.) This appears to be an advantage of hidden Markov models, as compared with some other time-series

Table 3: Comparison of models for gold particles data of Jung and Tremayne (2006a).

Model	$-l$	k	AIC	BIC
(Indep. observations on) a single Poisson	596.8215	1	1196	1200
Single CMP	596.7572	2	1198	1205
Independent mixture of two Poissons	596.5998	3	1199	1211
Two-state Poisson–HMM	557.4618	4	1123	1139
Two-state CMP–HMM	547.2147	6	1106	1130

models. We assume here that the missingness is non-informative. But if, for instance, the sample ACF is needed, the absence of some observations from the series is a significant complication.

8.2 Models for the 380 particle counts of Jung and Tremayne (2006a)

We report in Table 3 the likelihood, AIC and BIC values for five models fitted to the 380 counts, but report also in Section 8.3 on a wider range of models fitted to the fuller series. In this table is noticeable that the first three models (all of which assume independence) are by AIC and BIC clearly inferior to those allowing for serial dependence. This is to be expected. The two-state CMP–HMM has mean and variance 1.59 and 1.66, so it is indeed slightly overdispersed, as is the sample (cf. 1.56 and 1.62). The model has $\rho(k) = 0.3143 \times 0.9314^k$ and $\rho(1) = 0.29$, which is very much smaller than the sample first-order autocorrelation (0.57). We shall (in Section 8.3) see that a two-state CMP–HMM for the full series behaves very similarly, both as regards the overdispersion and the first-order autocorrelation.

8.3 Models for the 1598 particle counts of Guttorp (1991)

In this section we report on a range of models fitted to the series as given by Guttorp, and on just one model (a two-state CMP–HMM) fitted to the original series as given by Westgren (1916, Versuchsreihe C), i.e. with those 11 observations treated as missing and

Table 4: Two-state CMP–HMMs for gold particles data of Guttorp. Models were fitted both with and without imputation of the missing items. Figures appearing in smaller type below parameter estimates are approximate standard errors found from the numerical Hessian supplied by `nlm`.

Parameter	Estimated with imputation	Estimated without imputation
$\boldsymbol{\lambda}$	(1.396, 10.97)	(1.403, 11.56)
	0.102, 3.27	0.102, 3.50
$\boldsymbol{\nu}$	(2.358, 2.257)	(2.343, 2.288)
	0.184, 0.221	0.180, 0.225
$\boldsymbol{\Gamma}$	$\begin{pmatrix} 0.9569 & 0.0431 \\ 0.0832 & 0.9168 \end{pmatrix}$	$\begin{pmatrix} 0.9570 & 0.0430 \\ 0.0843 & 0.9157 \end{pmatrix}$
	$\begin{pmatrix} & 0.0089 \\ 0.0167 & \end{pmatrix}$	$\begin{pmatrix} & 0.0088 \\ 0.0165 & \end{pmatrix}$
$\boldsymbol{\delta}$	(0.6585, 0.3415)	(0.6619, 0.3381)
–log-likelihood	2141.018	2128.438
model mean	1.421	1.425
sample mean	1.425	1.428
model variance	1.499	1.505
sample variance	1.508	1.514
model ACF, $\rho(k)$	0.4791×0.8737^k	0.4830×0.8727^k
sample $\rho(1)$	0.64	

Table 5: Comparison of models for gold particles data as given by Guttorp (1991, pp. 191–192), i.e. with 11 missing items imputed.

Model	$-l$	k	AIC	BIC
(Indep. observations on) a single Poisson	2433.203	1	4868	4874
Single CMP	2432.470	2	4869	4880
Independent mixture of two Poissons	2433.203	3	4872	4889
Two-state Poisson–HMM	2221.281	4	4451	4472
Two-state CMP–HMM	2141.018	6	4294	4326
Three-state CMP–HMM	2047.218	12	4118	4183
Four-state CMP–HMM	1998.432	20	4037	4144
Modified four-state CMP–HMM	1998.432	19	4035	4137

not imputed at all. This will be useful in assessing what effect the imputation has on the quantities of interest.

For Guttorp's series, all the models which assume independence are as usual inferior to the two-state Poisson–HMM. The latter has mean and variance 1.42 and 1.99 and ACF 0.2874×0.9430^k, hence $\rho(1) = 0.27$, compared to figures of 1.42, 1.51 and 0.64 for the sample. So the model overestimates the (very slight) overdispersion and underestimates the serial dependence, at least as represented by the first-order autocorrelation. When confronted with the combination of very slight overdispersion and substantial serial dependence, the Poisson–HMM steers a middle course.

We give the two two-state HMMs with CMP state-dependent distributions in full in Table 4, along with approximate standard errors, and report more briefly on the others in Table 5. Table 4 displays the slight overdispersion in the models, as in the sample. Table 4 also displays both the (typically small) effect of the imputation on the MLEs and standard deviations, and its very small effect on the model mean, variance and ACF. Included in Table 5 are CMP–HMMs with three and four states, which according to AIC and BIC are far better than the two-state model, in spite of having 12 and 20 parameters respectively. At first sight this is surprising, but a comparison of expected with actual

Table 6: Gold particles data of Guttorp, comparison of actual frequencies with those expected under CMP–HMMs with 2–4 states.

	0	1	2	3	4	5	6	7	8	9
Expected (2 states)	380.7	600.7	324.3	182.5	81.3	23.4	4.5	0.6	0.1	0.0
Expected (3)	380.2	588.7	355.2	162.0	78.0	27.8	5.5	0.6	0.0	0.0
Expected (4)	383.3	576.3	367.6	162.6	73.1	28.4	6.0	0.7	0.0	0.0
Actual	384	575	361	176	67	28	5	2	0	0

Table 7: Gold particles data of Guttorp, comparison of sample moments with those of CMP–HMMs with 2–4 states.

	mean	variance	$\rho(1)$	$\rho(2)$	$\rho(20)$	AIC	BIC
Two-state model	1.421	1.499	0.42	0.37	0.032	4294	4326
Three-state	1.423	1.507	0.55	0.47	0.024	4118	4183
Four-state	1.423	1.504	0.60	0.50	0.017	4037	4144
Sample	1.425	1.508	0.64	0.53	0.193		

frequencies goes some way toward explaining why the two-state model is so inferior; see Table 6. Notice especially the expectations of 600.7 and 324.3 under the two-state model, which differ markedly from the corresponding actual frequencies. However, none of the three CMP–HMMs fits the ACF particularly well; see Table 7. In all three cases the model ACF approaches zero far quicker than does the sample ACF.

In view of Tables 6 and 7, it is plausible that the models which are favored by AIC and BIC are those which do a better job of matching the marginal distribution and the autocorrelations at low lags. Higher-order autocorrelations seem less important, and the three models all do a good job of matching the sample mean and variance.

A clear difference between the three- and four-state models is this. In the former the three values of ν were all between 3 and 4. In the latter there was, as in Section 5.4, one very large value of ν, 31.4, although the exact value of ν is irrelevant once it exceeds a certain level, especially if the corresponding λ is small. In this case λ is very small, 0.0403.

Again, this represents essentially a Bernoulli distribution. The model-fitting process has created a state which allows for the values 0 and 1 only. We therefore fitted a modified four-state model, which is as follows. The TPM is

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.738 & 0.216 & 0.046 & 0.000 \\ 0.079 & 0.804 & 0.114 & 0.004 \\ 0.000 & 0.194 & 0.734 & 0.072 \\ 0.029 & 0.000 & 0.165 & 0.806 \end{pmatrix},$$

with corresponding $\boldsymbol{\delta} = (0.147, 0.447, 0.289, 0.117)$. In state 1 the observation has a Bernoulli distribution with probability 0.0387 of taking the value 1. In states 2–4 the observations have CMP distributions with parameters $\boldsymbol{\lambda} = (3.053, 119.0, 224.1)$ and $\boldsymbol{\nu} = (4.220, 5.286, 3.887)$. The maximized log-likelihood is -1998.432, as was the case for the unmodified four-state model. Because the estimates are on or very close to the boundary of the parameter space, we did not compute standard errors here.

To sum up: of the models considered, the simplest that can cope with both serial dependence and the very slight overdispersion is the two-state CMP–HMM. The corresponding three- and four-state models are much better at matching the low-order autocorrelations and the observed marginal frequencies, but have many more parameters. The four-state model can be slightly simplified by replacing one of the CMP distributions by a Bernoulli distribution.

9 Discussion

The R code used by us to fit CMP–HMMs relies on the routine `dcmp` provided by the package `COMPoissonReg`, used with default settings. This routine was tested against `dpois` for the Poisson special case, and more generally against our own code, with 1 001 terms used in our code to approximate the sum $Z(\lambda, \nu) = \sum_{k=0}^{\infty} \lambda^k/(k!)^\nu$. No material differences or difficulties were found. Such computations, if they misbehave, will tend to do so for the combination of large λ and small ν. There were some very large values of λ used in our CMP–HMMs, but the smallest ν used in such a model was 1.770. For $\nu = 1$ we tested `dcmp` for λ up to 50, and for $\nu = 2$, 3 and 4 we tested for λ up to 500. This covers the range of

interest in our CMP–HMMs, but in addition, for $\nu = 0.9$, we tested for λ up to 30.

Fortunately, such routines are, both in the applications reported here and more generally, needed more for $\nu > 1$ (underdispersion in the state-dependent distributions) than the case $\nu < 1$; if the counts being modeled are substantially overdispersed, a Poisson–HMM (or negative-binomial–HMM) would be our first port of call, not a CMP–HMM. But in our view caution is necessary if a small value of ν (below 1) is used in a CMP distribution with large λ. Shmueli *et al.* (2005, Appendix B) give an approximation for $Z(\lambda, \nu)$ which is especially accurate if $10^{\nu} < \lambda$, and least so when both λ and ν are small. The package `COMPoissonReg` uses truncation of the infinite series for Z rather than any such approximation, and the number of terms defaults to 100; if necessary, a larger number of terms can be used.

The data-sets of Sections 5–8 all display some serial dependence, but differ in their degree of underdispersion: those of Sections 5 and 6 are slightly underdispersed (mean-variance ratio about 0.94), that of Section 7 more so (ratio 0.73), and those of Section 8 slightly overdispersed (ratios 1.04 and 1.06). In all cases the two-state CMP–HMM was able to replicate the mean and variance closely while allowing for serial dependence, although the model ACF did not necessarily correspond closely with the sample ACF. This suggests that, if the extent of the under- or overdispersion in a series of counts is the main aspect of interest, a CMP–HMM with only two states will often be adequate.

Most of the HMMs reported in Sections 5–8 have two states, but models with three or more states can be considered for all the data-sets, and AIC or BIC used to select the number of states. Note, however, the warning of Pohle *et al.* (2017) against uncritical use of AIC and BIC to select the number of states in an HMM. Their extensive simulations suggest that, if one fits an overly simplistic HMM but then allows (only) the possibility of an increased number of states, any misspecification of the model will tend to result in extra states 'which, to some extent, absorb the neglected structure.' In these circumstances, even BIC (which penalizes extra parameters more than does AIC) tends to overestimate the number of states. Furthermore, the problem of multiple local optima seems likely to be greater for a larger number of parameters. Note that a large number of states in an HMM does not necessarily imply a large number of parameters. With suitable structuring

of the TPM it is possible for an HMM with (say) 50 or 100 states to have only three or four parameters. See for instance the HMMs used by Langrock *et al.* (2012) to fit stochastic volatility models.

It is clear that underdispersed distributions on the nonnegative integers other than the CMP could similarly be used as state-dependent distribution in at least one of the states. There may well be suitable distributions that are easier to compute with than is the CMP: the underdispersed distributions used by Weiß (2013) are possibilities. Essentially all that needs to be done to modify the computations is to replace the function evaluating the CMP probability mass function by one evaluating the desired distribution.

However, a further approach to time series of underdispersed counts is to use not a CMP but a distribution that is of finite support, e.g. a binomial. Effectively that is what was done in state 1 of the 'modified' model for pedestrian counts in Section 5.5, but a binomial could have been used there in both states if an upper bound to the number of pedestrians had been known; there was obviously some upper bound to the number of pedestrians that could fit into the block observed. There is available also a generalization of the binomial known as the Conway–Maxwell(–Poisson)-binomial distribution; see Shmueli *et al.* (2005, Sec. 2.4.2), Borges *et al.* (2014) and Kadane (2016). A further potentially useful distribution of finite support is a right-truncated Poisson; such a distribution can exhibit underdispersion. But our intention here is to draw attention to CMP–HMMs as a flexible and potentially useful class of models, not to explore in detail all the possible alternatives, nor to claim that CMP–HMMs are in some sense the best. We doubt whether we have said the last word on any of the data-sets discussed here.

Choosing among the various models that are possible for a given series of counts will often not be easy, but if the likelihood is available a comparison can be made by AIC and BIC, in addition to comparing sample and model quantities. If, however, one model has a plausible substantive interpretation and others do not, one can provisionally treat that model as preferable.

In conclusion: the class of stationary hidden Markov models with CMP distributions as state-dependent distributions is able to represent serial dependence plus under- or over-dispersion in a time series of counts, and — with or without some modification — can be

considered as a possibility for such series, although it is probably more likely to be useful in the underdispersed case. In all the applications discussed here, the models seem promising, but further experience of applications is needed to assess their practical utility.

Supplementary materials

All the data-sets discussed here accompany this paper, as does our R code to fit CMP–HMMs and other, simpler, models. The six data files are: `Fuerth_pedestrians_505.txt`, `Weiss_downloads_241.txt`, `Weiss_emergencies_96.txt`, `Jung_gold_380.txt`, `Guttorp_-gold_1598.txt`, and `soap.txt`. The three files of R code are: `AllFunctionsNeeded.R`, `CMPHMM_etc.R` and `CMPHMM_modified.R`.

Acknowledgements

Prof. Roland Langrock is thanked for his advice. Prof. Antonello Maruotti is thanked for his helpful suggestions on the computing in particular. Prof. Robert Jung is thanked for speedily providing the series of gold particle counts discussed in Section 8.2. The inter-library loans staff of the University of Cape Town are thanked for their excellent assistance. The James M. Kilts Center, University of Chicago Booth School of Business, is thanked for making available the soap sales data. The editor, the reviewers and (especially) the associate editor are thanked for their many suggestions, which have done much to improve this paper.

Appendix: Details of two results appearing in Section 2.2

Variance

Equation (1) for the variance seems difficult to find other than in Zucchini *et al.* (2016, p. 23). We give here some of the details of a proof. With the notation as before, the proof

starts from:

$$\mathrm{Var}(X_t) = \mathrm{E}(X_t^2) - (\mathrm{E}X_t)^2 = \sum_{i=1}^{m} \delta_i(\sigma_i^2 + \mu_i^2) - \left(\sum_{i=1}^{m} \delta_i \mu_i\right)^2,$$

and after that proceeds as follows.

$$\mathrm{Var}(X_t) = \sum_{i=1}^{m} \delta_i \sigma_i^2 + \text{`excess'},$$

where the excess term is given by

$$\sum_{i=1}^{m} \delta_i \mu_i^2 - \left(\sum_{i=1}^{m} \delta_i \mu_i\right)^2 = \sum_{i=1}^{m} \delta_i(1-\delta_i)\mu_i^2 - \sum_{i \neq j} \delta_i \delta_j \mu_i \mu_j = \sum_{i=1}^{m} \delta_i(1-\delta_i)\mu_i^2 - 2\sum_{i<j} \delta_i \delta_j \mu_i \mu_j.$$

Now apply

$$1 - \delta_i = \sum_{j \neq i} \delta_j = \sum_{j<i} \delta_j + \sum_{j>i} \delta_j,$$

and it follows that the excess is

$$\sum_{i<j} \delta_i \delta_j (\mu_j^2 + \mu_i^2 - 2\mu_i \mu_j) = \sum_{i<j} \delta_i \delta_j (\mu_i - \mu_j)^2,$$

as claimed. In another route (via the conditional variance formula), one of the terms is exactly $\sum_{i=1}^{m} \delta_i \sigma_i^2$, and the other is the excess as defined above.

Covariance

Equation (3) for the covariance appears, for instance, in Zucchini *et al.* (2016, p. 42), but we give here the key step of the proof. With notation as before, plus $\gamma_{ij}(k)$ denoting a k-step transition probability, that step is:

$$\mathrm{E}(X_t X_{t+k}) = \sum_{i=1}^{m} \sum_{j=1}^{m} \big(\delta_i \gamma_{ij}(k)\big)\big(\mu_i \mu_j\big) = \boldsymbol{\delta} \boldsymbol{M} \boldsymbol{\Gamma}^k \boldsymbol{\mu}'.$$

With the row vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as defined just after Equation (8), i.e. by $\boldsymbol{a} = \boldsymbol{\delta} \boldsymbol{M} \boldsymbol{U}$ and $\boldsymbol{b}' = \boldsymbol{U}^{-1} \boldsymbol{\mu}'$, that equation becomes

$$\mathrm{Cov}(X_t, X_{t+k}) = \boldsymbol{a}\, \mathrm{diag}\,(1, \omega_2^k, \ldots, \omega_m^k)\, \boldsymbol{b}' - (\boldsymbol{\delta} \boldsymbol{\mu}')^2.$$

Equation (9) then follows. Note that, if the first column of $\boldsymbol{U}$ is scaled so that it consists of ones, the first row of $\boldsymbol{U}^{-1}$ is $\boldsymbol{\delta}$, the stationary distribution. In that case, both a_1 and b_1 equal $\boldsymbol{\delta} \boldsymbol{\mu}'$.

References

Albert, P. S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, **47**, 1371–1381.

Borges, P., Rodrigues, J., Balakrishnan, N., and Bazan, J. (2014). A COM-Poisson type generalization of the binomial distribution and its properties and applications. *Statistics & Probability Letters*, **87**, 158–166.

Conway, R. W. and Maxwell, W. M. (1961). A queueing model with state dependent service rate. *The Journal of Industrial Engineering*, **XII**(2), 132–136.

Dunn, J. (2012). *compoisson: Conway-Maxwell-Poisson Distribution*. R package version 0.3.

Fürth, R. (1918). Statistik und Wahrscheinlichkeitsnachwirkung. *Physikalische Zeitschrift*, **XIX**, 421–426.

Fürth, R. (1919). Statistik und Wahrscheinlichkeitsnachwirkung: Nachtrag. *Physikalische Zeitschrift*, **XX**, 21.

Guttorp, P. (1991). *Statistical Inference for Branching Processes*. Wiley, New York.

Heyde, C. C. and Seneta, E. (1972). Estimation theory for growth and immigration rates in a multiplicative process. *Journal of Applied Probability*, **9**, 235–256.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, **17**(6), 359–380.

Jung, R. C. and Tremayne, A. (2006a). Coherent forecasting in integer time series models. *International Journal of Forecasting*, **22**, 223–238.

Jung, R. C. and Tremayne, A. R. (2006b). Binomial thinning models for integer time series. *Statistical Modelling*, **6**, 81–96.

Kadane, J. B. (2016). Sums of possibly associated Bernoulli variables: The Conway–Maxwell-binomial distribution. *Bayesian Analysis*, **11**(2), 403–420.

Langrock, R., MacDonald, I. L., and Zucchini, W. (2012). Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, **19**, 147–161.

Le, N. D., Leroux, B. G., and Puterman, M. L. (1992). Reader reaction: Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, **48**, 317–323.

MacDonald, I. L. (2017). Models for count data. *The American Statistician*, **71**(2), 187–190.

MacDonald, I. L. and Zucchini, W. (2016). Hidden Markov models for discrete-valued time series. In R. A. Davis, S. H. Holan, R. B. Lund, and N. Ravishanker, editors, *Handbook of Discrete-Valued Time Series*, pages 267–286. Chapman & Hall/CRC Press, London and Boca Raton, FL.

Martin, V. L., Tremayne, A. R., and Jung, R. C. (2014). Efficient method of moments estimators for integer time series models. *Journal of Time Series Analysis*, **35**, 491–516.

Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 270–293.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Safari, A., Altman, R. M., and Leroux, B. (2017). Parameter-driven models for time series of count data. *arXiv:1711.02753 [stat.ME]*.

Sellers, K., Lotze, T., and Raim, A. (2017). *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. R package version 0.4.1.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, **4**(2), 943–961.

Sellers, K. F. and Shmueli, G. (2013). Data dispersion: now you see it ... now you don't. *Communications in Statistics—Theory and Methods*, **42**(17), 3134–3147.

Sellers, K. F., Borle, S., and Shmueli, G. (2012). The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, **28**, 104–116.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **54**(1), 127–142.

Sur, P., Shmueli, G., Bose, S., and Dubey, P. (2015). Modeling bimodal discrete data using Conway-Maxwell-Poisson mixture models. *Journal of Business & Economic Statistics*, **33**(3), 352–365.

Weiß, C. H. (2007). Controlling correlated processes of Poisson counts. *Quality and Reliability Engineering International*, **23**, 741–754.

Weiß, C. H. (2008). Serial dependence and regression of Poisson INARMA models. *Journal of Statistical Planning and Inference*, **138**, 2975–2990.

Weiß, C. H. (2013). Integer-valued autoregressive models for counts showing underdispersion. *Journal of Applied Statistics*, **40**(9), 1931–1948.

Westgren, A. (1916). Die Veränderungsgeschwindigkeit der lokalen Teilchenkonzentration in kolloiden Systemen (Erste Mitteilung). *Arkiv för Matematik, Astronomi och Fysik*, **11**(14), 1–24.

Wu, G., Holan, S. H., and Wikle, C. K. (2013). Hierarchical Bayesian spatio-temporal Conway-Maxwell Poisson models with dynamic dispersion. *Journal of Agricultural, Biological and Environmental Statistics*, **18**(3), 335–356.

Yang, M., Cavanaugh, J. E., and Zamba, G. K. (2015). State-space models for count time series with excess zeros. *Statistical Modelling*, **15**(1), 70–90.

Zhu, F. (2012a). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications*, **389**, 58–71.

Zhu, F. (2012b). Modeling time series of counts with COM-Poisson INGARCH models. *Mathematical and Computer Modelling*, **56**(9-10), 191–203.

Zhu, L., Sellers, K. F., Morris, D. S., and Shmueli, G. (2017). Bridging the gap: A generalized stochastic process for count data. *The American Statistician*, **71**(1), 71–80.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**(1), 41–61.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Press, Boca Raton, FL, second edition.