Effective Bayesian Modeling of Groups of Related Count Time Series

Nicolas Chapados   CHAPADOS@APSTAT.COM

ApSTAT Technologies Inc., 408-4200 Boul. St-Laurent, Montral, QC, H2W 2R2, CANADA

Abstract

Time series of counts arise in a variety of forecasting applications, for which traditional models are generally inappropriate. This paper introduces a hierarchical Bayesian formulation applicable to count time series that can easily account for explanatory variables and share statistical strength across groups of related time series. We derive an efficient approximate inference technique, and illustrate its performance on a number of datasets from supply chain planning.

1. Introduction

Most classical time series forecasting models such as exponential smoothing (Hyndman et al., 2008) and ARIMA models (Box et al., 2008) assume that observations are real-valued and can take on both positive and negative values. In addition, the majority of classical approaches provide normal predictive distributions, if they do so at all. However, large segments of the practice of forecasting—for instance in supply chain planning—deal with time series that depart significantly from these assumptions: series, for example, that may consist only of non-negative integer observations, contain a large fraction of zeros, or are further characterized by long runs of zeros interspersed by some large non-zero values. In other words, the classical assumptions of conditional normality are grossly violated. Moreover, if multiple contemporaneous series are considered, common models either treat them completely independently, or—as in vector autoregressive models (Box et al., 2008)—attempt a more complex multivariate modeling that captures short-range cross-correlations but becomes unwieldy when managing hundreds of series; the common scenario of "weak coupling" between related series (e.g. consumer demand for a seasonal product at several stores of the same chain in a given city, which could share seasonal behavior but not strong cross-correlations) is not easily handled in classical modeling frameworks.

1.1. Motivating Applications

The starting point for the present work lies in the intermittent demand series that frequently occur in supply chain operations: these arise, for example, in the demand for spare parts in aviation, or in the number of "slow-moving" items sold in retail stores (Altay & Litteral, 2011). In addition to the non-negativity, integrality, skewness and high fraction of zeros attributes already outlined, these time series are commonly quite short: many weekly and monthly demand series encountered in practice may consist of some 30 to 100 observations. This makes it crucial to allow some information sharing across related series to more reliably capture posited common effects such as seasonalities and the impact of causal determinants (such as the market response to promotions or supply chain disruptions). These are the modeling challenges that we address in this paper.

1.2. Related Work

Most of the literature on intermittent demand forecasting relies on relatively simple techniques, typically variants of Croston's (1972) method which computes the expected demand in the next period as the ratio of the expected non-zero demand to the expected non-zero-demand time interval, both estimated by simple exponential smoothing. Croston's method produces point forecasts only; Shenstone & Hyndman (2005) studied variants with a proper stochastic foundation that can produce predictive intervals, although with no attempt to capture some known stylized facts of intermittent demand patterns such as heavy tails (Kwan, 1991). It is only with the recent work of Snyder et al. (2012) that a reasonably modern formulation was proposed in terms of a state-space model and distributional forecasts. This model still tracks the expected demand through an exponential-smoothing update, but emits predictive distributions that belong to the negative binomial family (described in the next section); model parameters are estimated by maximum likelihood (ML). Despite the evidence of improved accuracy against common benchmarks, this approach still exhibits a number of shortcomings: it is fundamentally univariate, does not easily allow explanatory variables, and the ML estimation framework does not reflect model parameter uncertainty that arises

with the very short series that are common in practice.

1.3. Contributions

This paper makes the following three contributions: (*i*) introducing a hierarchical probabilistic state-space model that is a good match to commonly-seen types of count data, allowing for explanatory variables and permitting information sharing across groups of related series (§2); (*ii*) introducing an effective inference algorithm for computing posterior distributions over latent variables and predictive distributions (§3); (*iii*) assessing the proposed approach's performance via a thorough experimental evaluation (§4).

2. Hierarchical Model for Non-Negative Integer Time Series

We introduce the proposed model in several stages, starting with the basic state-space structure and integer-valued observations (§2.1), introducing explanatory variables and structural zeros (§2.2), and finishing with the hierarchical structure allowing information sharing across series (§2.3).

2.1. Core Model

For a single non-negative time series, the model is expressed in state-space form (e.g., Durbin & Koopman, 2012), where the latent state η_t at period $t = 1, \ldots, T$ represents the log-expected value of the non-negative integer observation y_t. Of those, we shall assume that the first $T - h$ ($h \geq 1$) are observed, and the last h constitute the future values over which we would like to forecast. Representation in log-space enforces the constraint that the process mean can never become negative. The state space structure makes all observations independent of each other conditionally on the latent state; here we assume that observation y_t is drawn from a negative binomial (NB) distribution with mean $\exp \eta_t$ and size parameter α (which is independent of t),

$$y_t \sim \text{NB}(\exp \eta_t, \alpha).$$

The negative binomial distribution (parametrized by the mean μ instead of the more usual probability of success in a trial; e.g. Hilbe 2011) is given by

$$P_{\text{NB}}(y \mid \mu, \alpha) = \binom{\alpha + y - 1}{\alpha - 1} \left(\frac{\alpha}{\mu + \alpha}\right)^\alpha \left(\frac{\mu}{\mu + \alpha}\right)^y,$$

(1)

where $\binom{n}{m} \equiv \frac{\Gamma(n+1)}{\Gamma(m+1)\Gamma(n-m+1)}$ is the binomial coefficient. The negative binomial is appropriate for count data that is overdispersed with respect to a Poisson distribution (with the variance greater than the mean); the size parameter $\alpha > 0$ governs the level of overdispersion. The limiting case $\alpha \to \infty$ converges to a Poisson distribution.
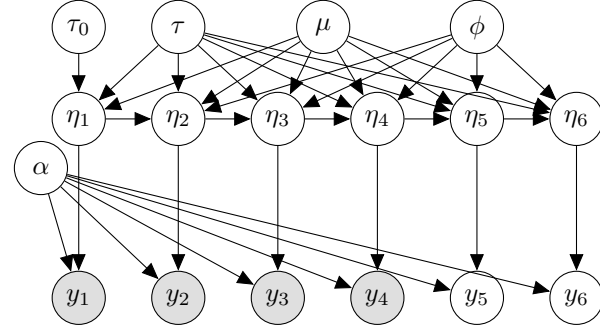

Figure 1. Basic state-space model for a single time series as a fully unfolded directed graphical model. Shaded nodes $\{y_1, \ldots, y_4\}$ are observed values; variables y_5 and y_6 are values to be forecasted. The dependence of the latent log-intensity process $\{\eta_t\}$ on all hyperparameters is made explicit.

The dynamics of the process log-mean η_t depend on the properties of the time series being modeled. For a stationary series, a mean-reverting autoregressive process is a sensible and tractable choice. In a supply chain context, mean reversion intuitively means that the long-run expected demand for an item, when projected far in the future, should fall back to a constant level in spite of any past transient disturbances. We express latent dynamics as an AR(1) process with normal innovations, with

$$
\begin{aligned}
\eta_1 &= \mu + \epsilon_1, \\
\eta_t &= \mu + \phi(\eta_{t-1} - \mu) + \epsilon_t, \quad t > 1, \\
\epsilon_1 &\sim \mathcal{N}(0, 1/\tau_0 + 1/\tau), \\
\epsilon_t &\sim \mathcal{N}(0, 1/\tau), \qquad\qquad t > 1,
\end{aligned}
$$

(2)

where $\mu \in \mathbb{R}$ is the long-run level of mean reversion, $-1 < \phi < 1$ is the speed of mean reversion, $\tau > 0$ is the precision of the process innovations, and $\tau_0 > 0$ allows for additional variance in the initial period. All ϵ_t are assumed mutually independent. The model structure is depicted in graphical form in Fig. 1. Forecasting in this model conceptually proceeds in three steps: (*i*) from the observed values of the time series, we carry out inference over all unobserved model variables (the clear nodes in Fig. 1); (*ii*) using the inferred process parameters (τ, μ, ϕ), we project the latent dynamics into the future to obtain a distribution over future values of the latent state (η_5 and η_6 in the figure); and (*iii*) obtaining a predictive distribution over future observations (y_5 and y_6 in the figure). This description can easily be extended to accommodate multivariate observations and latent states; the latent states would then follow a vector autoregressive (VAR) process.

2.2. Explanatory Variables and Structural Zeros

Explanatory variables (which can include seasonal terms, as well as factors that causally impact the observed time series) can be incorporated by viewing them as *local forcing terms* that temporarily shift the location of the latent
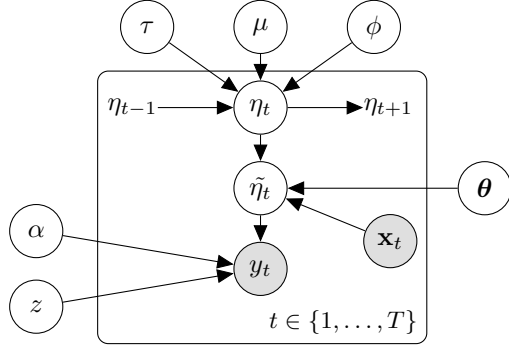
Figure 2. Incorporating explanatory variables $\mathbf{x}_t$ and zero-inflation z into the model, where the plate indicates repetition over time periods t. For brevity, license is taken to omit depiction of the initial and final latent log-intensity, as well as representing the unshaded $\{y_t\}$ over the forecasting period.

process mean. This is illustrated in Fig. 2. We assume that explanatory variables at period t, $\mathbf{x}_t \in \mathbb{R}^N$, are always observed, non-stochastic and known ahead of time (so that we know the future values of $\{\mathbf{x}_t\}$ over the forecasting horizon). They are linearly combined through regression coefficients $\boldsymbol{\theta}$ to additively shift the latent η_t, yielding an effective log-mean $\tilde{\eta}_t$,

$$\tilde{\eta}_t = \eta_t + \mathbf{x}_t'\boldsymbol{\theta}, \tag{3}$$

where the $\{\eta_t\}$ follows the same AR(1) process as previously. In latent (log) space, the addition operation corresponds to a multiplicative impact of explanatory variables on the process mean in observation space, which is often a good fit to the underlying data generating process (e.g. seasonalities, or consumer response to promotions or special events). It also makes the regression coefficients $\boldsymbol{\theta}$ relatively independent of the scale of the series, and makes it easier to share information across multiple time series as described in §2.3.

In many real-world series, one observes an excess of zero values compared to the probability under the NB (e.g., Lambert, 1992): this can arise for structural reasons in the underlying process (e.g., out-of-stock items or supply chain disruptions in a retail store context, both of which would override the natural consumer demand modeled by the NB). For these reasons, we add extra unconditional mass at zero, yielding so-called *zero-inflated* NB observations,

$$y_t \sim z\,\delta_0 + (1-z)\,\mathrm{NB}(\exp\tilde{\eta}_t, \alpha), \tag{4}$$

where δ_0 represents unit probability mass at zero and $z \in [0,1]$ is the probability of structural zero. We assume a $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ prior for z. This is our final observation model.

2.3. Sharing Information Across Multiple Time Series

Finally, we allow for a group of L related time series to share information, particularly in the form of a shared
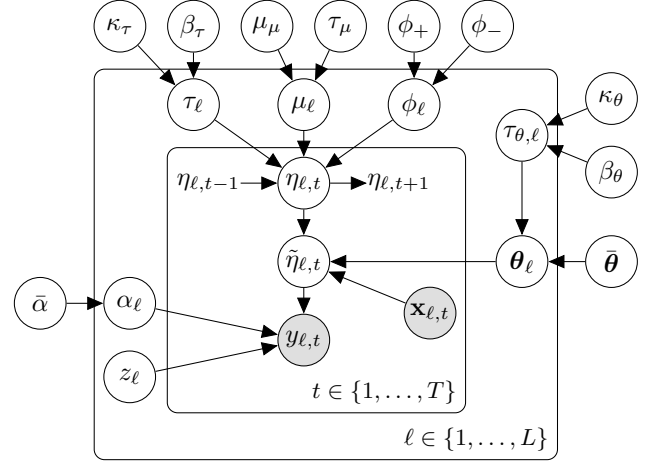

Figure 3. Model with hierarchical structure, where the outer plate indicates repetition across several time series ℓ. Information sharing across series is achieved by "global" hyperparameters located outside all plates, in particular the regression coefficient prior $\bar{\boldsymbol{\theta}}$.

hyperprior over regression coefficients and latent process characteristics. In the spirit of hierarchical models studied in statistics (Gelman & Hill, 2007) and machine learning (Teh & Jordan, 2010; Fox et al., 2010), we let those parameters (for all time series $\ell \in \{1, \ldots, L\}$ that belong to the group being modeled simultaneously) share common parents, as illustrated on Fig. 3. A new plate iterates over the series-level parameters, which inherit as follows from "global" parameters shared across all time series:

$$\begin{aligned}
\alpha_\ell &\sim \mathrm{Exponential}(\bar{\alpha}), & \mu_\ell &\sim \mathcal{N}(\mu_\mu, 1/\tau_\mu), \\
\tau_\ell &\sim \Gamma(\kappa_\tau, \beta_\tau), & \tau_{0,\ell} &\sim \Gamma(\kappa_{0,\tau}, \beta_{0,\tau}), \\
\phi_\ell &\sim \mathrm{Beta}(\phi_+ + \phi_-, \phi_-), & \boldsymbol{\theta}_\ell &\sim \mathcal{N}(\bar{\boldsymbol{\theta}}, \tfrac{1}{\tau_{\theta,\ell}}I), \\
\tau_{\theta,\ell} &\sim \Gamma(\kappa_\theta, \beta_\theta),
\end{aligned}$$

where $\Gamma(a, b)$ represents the gamma distribution with shape parameter a and scale parameter b, and $\mathrm{Beta}(a, b)$ is the beta distribution with shape parameters a and b. The series-level parameters (variables in plate ℓ on Fig. 3) all have the same meaning as previously, except that the ℓ index makes them dependent on a specific time series. The hyperpriors that are used for the global parameters are given in the supplementary material. The latent dynamics of $\{\eta_{\ell,t}\}$ and the observation model are the same as in the previous sections. For convenience, we shall denote all "global" variables in Fig. 3 (except μ_μ, for reasons to be made clear shortly) by $\boldsymbol{\Theta}_G = \{\bar{\alpha}, \tau_\mu, \kappa_\tau, \beta_\tau, \kappa_{0,\tau}, \beta_{0,\tau}, \kappa_\theta, \beta_\theta, \phi_+, \phi_-, \bar{\boldsymbol{\theta}}\}$, all series-$\ell$-level variables (except μ_ℓ) by $\boldsymbol{\Theta}_\ell = \{z_l, \alpha_\ell, \tau_\ell, \tau_{0,\ell}, \phi_\ell, \boldsymbol{\theta}_\ell, \tau_{\theta,\ell}\}$, and all latents over which we should do inference by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_G \cup \{\boldsymbol{\Theta}_\ell\} \cup \{\mu_\mu, \mu_\ell, \eta_{\ell,t}\}$.

In the remainder of this paper, we call this model the *hierarchical negative-binomial state space* (**H-NBSS**) model. It must be stressed that this model assumes that all time series in the group are conditionally independent given

the series-level parameters. In particular, the model does not allow expressing observation-level cross-correlations across different time series $\ell_i \neq \ell_j$, except through common effects coming from explanatory variables.[1]

3. Inference

Due to the non-conjugacy between the zero-inflated negative-binomial likelihood and the normal latent log-mean process prior (and the general difficulty of finding useful conjugate priors for negative-binomial likelihoods), inference in the H-NBSS model does not have an analytically tractable solution. One must resort to approximation techniques, which fall, broadly speaking, into two families: deterministic and stochastic methods (Barber et al., 2011). Of the deterministic approaches, early examples include assumed density filtering (ADF, Maybeck, 1979), which is a sequential projection approach, as well as numerical integration schemes such as the piecewise approximation of Kitagawa (1987). More recently, the expectation propagation (EP) algorithm of Minka (2001), a generalization of ADF, has proved successful in a number of non-linear filtering and smoothing problems (Heskes & Zoeter, 2002; Yu et al., 2006; Deisenroth & Mohamed, 2012). As to stochastic approaches, they can take the form of variants of Gibbs sampling, such as the recursive forward-filtering backward-sampling (FFBS) algorithm (Robert et al., 1999; Scott, 2002) as well as sequential Monte Carlo techniques such as particle filtering (reviewed by Doucet et al., 2001). Durbin & Koopman (2000) present an alternative approach based on importance sampling and antithetic variables. Static hierarchical regression models have been widely studied in the statistics literature (Gelman & Hill, 2007), where typical inference techniques rely on block Gibbs sampling. Dynamic hierarchical models have been less commonly studied, with the notable exception of the nonparametric Bayesian model of Fox et al. (2010), who use an efficient form of the Metropolis-Hastings algorithm for inference

It is imperative to contrast the benefits of a proposed algorithm to the requirements of forecasting practice: for instance, in a supply chain context, it is routine business to process tens to hundreds of millions of time series on a daily or weekly basis.[2] Despite their inadequacies, practitioners still rely on very computationally simple methods such as exponential smoothing for the vast majority of their tasks. Needless to say, for a forecasting approach to have an impact in practice, its accuracy benefits must justify its computational cost. This seems to rule out all stochastic

algorithms, as well as many deterministic ones such as EP.

We shall argue that for the H-NBSS model, a Gaussian approximation of the latent variables at their posterior mode, known as the Laplace approximation (Bishop, 2006), yields near-optimal performance at extremely attractive computational cost compared to the alternatives. One reason to expect good performance is that most of the important (for forecasting) latent variables in the model ($\{\mu_\mu, \mu_\ell, \eta_{\ell,t}\}$) have a conditionally normal prior; their posterior is nearly always close to normality despite the non-linear likelihood.

3.1. Posterior Calculation

The Laplace approximation requires to calculate the log-posterior probability up to an additive constant,

$$\log P(\boldsymbol{\Theta} \mid \mathbf{Y}) = \log P(\boldsymbol{\Theta}) + \log P(\mathbf{Y} \mid \boldsymbol{\Theta}) + C, \quad (5)$$

where $\mathbf{Y} = \{y_{\ell,t}\}_{t=1}^{T-h}$ is the set of all observed series values in all groups and C is an unknown (and for the Laplace approximation, unimportant) constant. The log-likelihood term $\log P(\mathbf{Y} \mid \boldsymbol{\Theta})$ is derived straightforwardly from the observation model (4) along with the negative binomial probability distribution (1). The first term—the log-prior—decomposes into global-, series- and observation-level terms,

$$\log P(\boldsymbol{\Theta}) = \log P(\boldsymbol{\Theta}_G) + \sum_{\ell=1}^{L} \log P(\boldsymbol{\Theta}_\ell \mid \boldsymbol{\Theta}_G) + \\ \log P(\mu_\mu) + \sum_{\ell=1}^{L} \log P\big(\mu_\ell, \{\eta_{\ell,t}\} \mid \boldsymbol{\Theta}_\ell, \mu_\mu\big).$$

The second line of this equation expresses a prior over jointly normally-distributed variables with a highly structured (and very sparse) precision matrix, a Gaussian Markov random field (GMRF, see Rue & Held, 2005), to which we now turn.

3.2. GMRF Prior

For the single time series process described in (2), assuming that the initial η_1 has the long-run process distribution with precision of $1/(\tau(1-\phi^2))$,[3] the joint prior over the latent process $\{\eta_t\}$ along with normally-distributed long-run mean μ (having prior precision τ_μ) is normally distributed with a *tridiagonal precision matrix* Q, except for the last row and column (corresponding to μ), that follows the pattern

$$Q = \begin{bmatrix} \tau & -\tau\phi & 0 & 0 & 0 & \tau\tilde{\phi} \\ -\tau\phi & \tau(\phi^2+1) & \ddots & 0 & 0 & -\tau\tilde{\phi}^2 \\ 0 & -\tau\phi & \ddots & -\tau\phi & 0 & \vdots \\ 0 & 0 & \ddots & \tau(\phi^2+1) & -\tau\phi & -\tau\tilde{\phi}^2 \\ 0 & 0 & 0 & -\tau\phi & \tau & \tau\tilde{\phi} \\ \tau\tilde{\phi} & -\tau\tilde{\phi}^2 & \cdots & -\tau\tilde{\phi}^2 & \tau\tilde{\phi} & \tau_\mu + \tau\psi_T \end{bmatrix},$$

---

[1] And except, of course, if the observations $y_{\ell,t}$ are themselves multivariate, which is outside the scope of this paper.

[2] For example, a large department retail store may sell 100K different items (Stock Keeping Units, SKUs); a chain with 1000 stores would then require periodic forecasts for 100M series.

[3] So that τ_0 would equal $\phi^2/(\tau(1-\phi^2))$.

where $\tilde{\phi} \equiv \phi - 1$ and $\psi_T \equiv T - 2(T-1)\phi + (T-2)\phi^2$, where T is the number of observations. The determinant of this matrix is $\tau^T \tau_\mu (1 - \phi^2)$.[4] The sparsity of Q makes it extremely fast to compute the process prior term.

When considering the hierarchical model of section 2.3, a similar sparsity pattern holds: the joint precision matrix across all variables that belong to the GMRF prior has block-diagonal structure, with one sub-matrix like Q for each time series, and a final row/column linking the series means $\{\mu_\ell\}$ to the global mean μ_μ. Details are given in the supplementary material.

3.3. Optimization and Predictive Distribution

Maximization of (5) over Θ can be carried out efficiently using Quasi-Newton methods such as L-BFGS (Nocedal & Wright, 1999).[5] Let $\hat{\Theta}$ be the maximizing value and $\hat{\Sigma}$ the inverse of $H_{\hat{\Theta}}$, the Hessian matrix of (5) evaluated at $\hat{\Theta}$. The Laplace approximation posits that the posterior distribution over Θ is jointly normal with mean $\hat{\Theta}$ and covariance matrix $\hat{\Sigma}$. Due to the structure of the GMRF prior, matrix $H_{\hat{\Theta}}$ is nearly block-diagonal except for the variables that belong to Θ_G. This makes it efficient to compute $\hat{\Sigma}$ by sparse matrix solvers (e.g., Davis, 2006).

From the graphical model structure (Fig. 3) and the observation model (4), the predictive distribution over a future value $y_{\ell,t}, t \geq T - h + 1$ depends only on the distributions of $\tilde{\eta}_{\ell,t}, \alpha_\ell$ and z_ℓ. In practice, the posterior uncertainty over z_ℓ and α_ℓ is small and can be neglected. From the observation model, the posterior distribution over $y_{\ell,t}$ can be obtained by integrating out $\tilde{\eta}_{\ell,t}$,

$$P(y_{\ell,t} \mid \mathbf{Y}) = \int_{-\infty}^{\infty} P(y_{\ell,t} \mid \exp \tilde{\eta}_{\ell,t}) \, P(\tilde{\eta}_{\ell,t} \mid \mathbf{Y}) \, d\tilde{\eta}_{\ell,t}$$
$$= P\Big(y_{\ell,t} \mid \int_{-\infty}^{\infty} \exp \tilde{\eta}_{\ell,t} P(\tilde{\eta}_{\ell,t} \mid \mathbf{Y}) \, d\tilde{\eta}_{\ell,t}\Big),$$

where a key use of the well-known summation property of the negative binomial is made, wherein for IID variables $X_i \sim \mathrm{NB}(\mu_i, \alpha)$, we have $\sum_i X_i \sim \mathrm{NB}\big(\sum_i \mu_i, \alpha\big)$, and we assume that the summation converges to an integral in the limit. Hence, only the posterior expectation of $\exp \tilde{\eta}_{\ell,t}$ is needed, which is readily obtained as

$$\mathbb{E}\big[\exp \tilde{\eta}_{\ell,t} \mid \mathbf{Y}\big] = \exp\Big(\mathbb{E}\left[\tilde{\eta}_{\ell,t} \mid \mathbf{Y}\right] + \tfrac{1}{2}\mathrm{Var}\left[\tilde{\eta}_{\ell,t} \mid \mathbf{Y}\right]\Big),$$

---

[4]These results were obtained by direct symbolic matrix inversion in Mathematica, and can be verified to yield the identity matrix when multiplying Q with the joint $\{\eta_t, \mu\}$ covariance matrix.

[5]Local optima are not a problem in practice as long as the $\{\eta_{\ell,t}\}$ are suitably initialized; a reasonable initialization for $\eta_{\ell,t}, 1 \leq t \leq T - h$ can be taken as the midpoint between $\log y_{\ell,t}$ and the mean of log-values for series ℓ, m_ℓ. Initial values for $T - h + 1 \leq t \leq T$ can be taken to be m_ℓ.

since $\tilde{\eta}_{\ell,t}$ has a normal posterior under the Laplace approximation. From (3),

$$\mathbb{E}\left[\tilde{\eta}_{\ell,t} \mid \mathbf{Y}\right] = \mathbb{E}[\eta_{\ell,t} \mid \mathbf{Y}] + \mathbf{x}_{\ell,t}'\mathbb{E}[\boldsymbol{\theta}_\ell \mid \mathbf{Y}],$$

where the conditional posteriors for $\eta_{\ell,t}$ and $\boldsymbol{\theta}_\ell$ are directly available in $\hat{\Theta}$. Similarly, the posterior variance for $\tilde{\eta}_{\ell,t}$ is

$$\mathrm{Var}\left[\tilde{\eta}_{\ell,t} \mid \mathbf{Y}\right] = \mathrm{Var}[\eta_{\ell,t} \mid \mathbf{Y}] +$$
$$\mathbf{x}_{\ell,t}'\mathrm{Var}[\boldsymbol{\theta}_\ell \mid \mathbf{Y}]\mathbf{x}_{\ell,t} + 2\mathbf{x}_{\ell,t}'\mathrm{Cov}[\eta_{\ell,t}, \boldsymbol{\theta}_\ell \mid \mathbf{Y}]$$

where the variances and covariances on the right-hand side are from $\hat{\Sigma}$.

3.4. Accuracy Compared to MCMC

Ultimately, the validity of approximate inference is predicated on its empirical performance, which is evaluated in the next section. Here, we graphically contrast on Fig. 4 the inference results for a single series between the Laplace approximation outlined previously and an equivalent model computed with Markov chain Monte Carlo (MCMC). The latter is implemented in the Stan modeling language (Stan Development Team, 2013), which uses the "no-U-turn" variant (Hoffman & Gelman, 2013) of Hamiltonian Monte Carlo (HMC). We combined the results of four independent chains, each run with 1500 burn-in iterations followed by 18500 sampling iterations. Overall, we note the similarity of the posterior distributions between the two approaches, although the Laplace approximation slightly underestimates the posterior variance in η_t over the forecast horizon (the region denoted "Fcast" in the plots) compared with MCMC. Should additional accuracy be required in the Laplace approximation, one could turn to a numerical integration technique for hyperparameters of models equipped with GMRF priors, the so-called integrated nested Laplace approximation (INLA, Rue et al., 2009).

Of significant importance for practical applications, however, is computational time. For the results illustrated in Fig. 4, whereas our implementation of the Laplace approximation (coded in the interpreted language R) converges in a few seconds, a roughly equivalent run of the MCMC sampler takes 30 times longer (and Stan is a very efficient engine, compiling the model into C++ code with analytical gradient computation for HMC). As will be clear from the experimental results, the additional computational cost of MCMC does not translate into a performance advantage in forecasting.

4. Experimental Evaluation

We evaluate model performance on three datasets obtained from supply chain operations. The first one (RAID) is the sales of bug spray at 26 locations of a major US retailer.
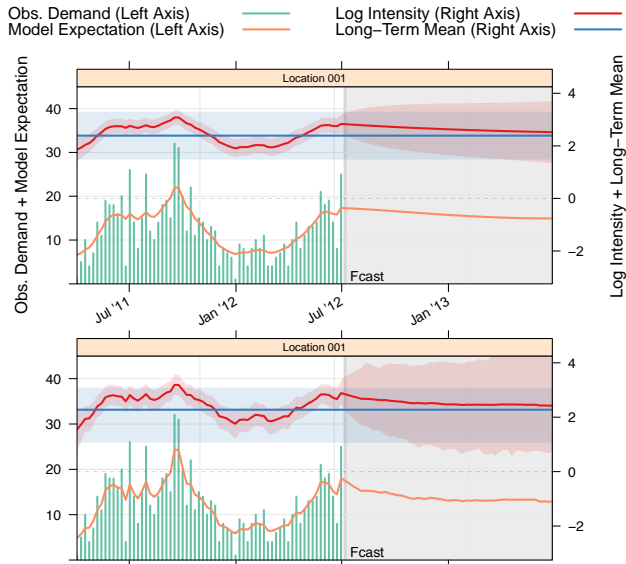
Figure 4. Comparison of inference results between the **Laplace approximation (top)** and **MCMC (bottom)**. The observed demand (green vertical lines) and model expectation of y_t (continuous orange line) should be interpreted according to the **left axis**. The latent variables η_t (red curve, denoted "log-intensity") and level of mean-reversison (blue horizontal line) should be interpreted according to the **right axis**. Shaded areas are 95% credible intervals. The posteriors between the two approaches are close.

The second one (GLUE) is the sales of gluestick at 2033 US retail locations. The third (PARTS) is the demand for spare parts of a major European IT firm, previously studied by Syntetos et al. (2012). The first two datasets illustrate variability at the location level for the same item (SKU), whereas the third one illustrates variability at the item level. All time series are non-negative integers, some with a large fraction of zeros, and all series of a given dataset covers the same date range; Table 1 shows some summary statistics. The "mean non-zero value" is the mean series value, conditional on the value being positive; the "mean non-zero inter-period" is the number of periods between non-zero observations; and the "mean sq. coef. of variation" is the square of the coefficient of variation (CV^2), which is σ^2/m^2 for a series with mean m and standard deviation σ.

Model performance is evaluated by the out-of-sample forecasting accuracy over the horizon h (where h varies from 1 to 12 periods), measured according to the negative log-likelihood (NLL) per period, relative mean squared error (MSE) and relative mean absolute error (MAE). To reduce the scale dependence of the MSE and MAE, the MSE is normalized by the in-sample variance and the MAE is normalized by the in-sample mean absolute deviation from the in-sample mean. We evaluate performance by a sequential re-training procedure that alternates between model training and testing, moving at each iteration the first observation of the (previous) test set to the end of the (new) train-

Table 1. Summary statistics of datasets evaluated. The categorization of demand patterns into "smooth" (high non-zero demand rate, low CV^2), "erratic" (both high), "intermittent" (both low) and "lumpy" (low demand rate, high CV^2) follows Syntetos et al. (2005). The chosen datasets cover a broad mix of patterns.

	RAID	GLUE	PARTS
Number of time series	24	2033	3055
Sampling period	Weekly	Weekly	Monthly
Nb. of observations per series	66	79	48
Initial training set duration	53	65	24
Mean non-zero value	6.31	1.53	23.73
Mean non-zero inter-period	1.15	4.02	3.82
Mean sq. coef. of variation (CV^2)	0.39	0.29	1.29
% Smooth	57%	1%	4%
% Erratic	27%	0%	15%
% Intermittent	17%	90%	14%
% Lumpy	0%	9%	67%

ing set. This simulates the action of a decision-maker acting in real-time, retraining models as new information becomes available. All reported results are averages of out-of-sample performance under this procedure. The initial training set durations for each dataset are given in Table 1.

4.1. Benchmark Models

We compare the forecasting performance of the proposed H-NBSS model against the following benchmarks: (*i*) Croston's (1972) method, (*ii*) simple exponential smoothing (E-S) with additive errors and an automatically-adjusted smoothing constant, and (*iii*) the damped dynamic model with negative binomial observations of Snyder et al. (2012). The first two approaches are as implemented by the corresponding functions in the R `forecast` package (Hyndman et al., 2013). Since they provide point forecasts only, we evaluate predictive distributions under two alternatives: a Gaussian distribution, with variance given by the variance of training residuals, or a Poisson distribution. In both cases, the mean is given by the point forecast.

4.2. Performance Results

Seasonalities are significant on the RAID dataset, which are incorporated into the H-NBSS through explanatory variables; for this dataset, we report H-NBSS results both without and with seasonalities, for both the Laplace and MCMC approximate inference. Seasonalities are omitted from GLUE and PARTS for space reasons since they yield very similar performance to the model without seasonal effects. Moreover, H-NBSS results in this section consider each dataset series independently of the others (i.e. groups of size 1). The benefits of hierarchy are examined in the next section. Out-of-sample performance results for all datasets and models at selected forecasting horizons are given in Table 2. NLL results at all horizons appear in Fig. 5. We observe that on the NLL measure (which

Table 2. Forecasting performance at various horizons. For all measures, a lower value indicates a higher accuracy; best results are bolded.

|	Normalized NLL			Relative MSE			Relative MAE		
	Dataset: RAID								
Forecast Horizon (periods)	1	4	8	1	4	8	1	4	8
Croston (Gaussian)	2.408	2.461	2.523	0.879	1.003	1.133	0.900	0.969	1.024
Croston (Poisson)	2.457	2.620	2.827	0.879	1.003	1.133	0.900	0.969	1.024
E-S Additive (Gaussian)	2.284	2.356	2.492	0.708	0.830	1.047	0.824	0.883	0.981
E-S Additive (Poisson)	2.312	2.447	2.775	0.708	0.830	1.047	0.824	0.883	0.981
Snyder-Ord-Beaumont	2.285	2.340	2.384	0.749	0.822	0.869	0.836	0.871	0.893
H-NBSS w/o Seas (Laplace)	2.251	2.275	2.378	0.677	0.732	0.837	0.811	0.836	0.876
H-NBSS w/o Seas (MCMC)	2.286	2.380	2.595	0.728	0.847	1.026	0.835	0.893	0.972
H-NBSS w/ Seas (Laplace)	**2.228**	**2.219**	**2.265**	0.656	0.669	0.714	0.800	**0.808**	**0.814**
H-NBSS w/ Seas (MCMC)	2.231	2.278	2.362	**0.634**	**0.654**	**0.693**	**0.796**	0.821	0.849
	Dataset: GLUE								
Forecast Horizon (periods)	1	4	8	1	4	8	1	4	8
Croston (Gaussian)	1.238	1.256	1.269	1.126	1.141	1.144	1.029	1.032	1.031
Croston (Poisson)	0.881	0.880	0.880	1.126	1.141	1.144	1.029	1.032	1.031
E-S Additive (Gaussian)	1.241	1.254	1.267	1.123	1.132	1.130	0.997	0.998	0.994
E-S Additive (Poisson)	0.871	0.868	0.866	1.123	1.132	1.130	0.997	0.998	0.994
Snyder-Ord-Beaumont	0.872	0.876	0.889	1.122	1.137	1.139	0.982	0.985	0.985
H-NBSS w/o Seas (Laplace)	**0.830**	**0.827**	**0.825**	1.112	**1.124**	1.127	**0.975**	**0.976**	**0.977**
H-NBSS w/o Seas (MCMC)	0.835	0.832	0.829	**1.111**	**1.124**	**1.126**	1.002	1.006	1.006
	Dataset: PARTS								
Forecast Horizon (periods)	1	4	8	1	4	8	1	4	8
Croston (Gaussian)	68.28	78.87	98.66	**132.39**	**150.31**	**155.10**	1.982	2.212	2.332
Croston (Poisson)	14.53	15.88	16.76	**132.39**	**150.31**	**155.10**	1.982	2.212	2.332
E-S Additive (Gaussian)	69.07	78.11	81.19	132.50	150.45	155.26	1.879	2.116	2.238
E-S Additive (Poisson)	14.17	16.52	18.49	132.50	150.45	155.26	1.879	2.116	2.238
Snyder-Ord-Beaumont	8.80	19.48	38.24	132.51	150.48	155.28	**1.833**	**2.063**	**2.180**
H-NBSS w/o Seas (Laplace)	**4.21**	**4.51**	**4.91**	132.48	150.43	155.23	1.866	2.097	2.217
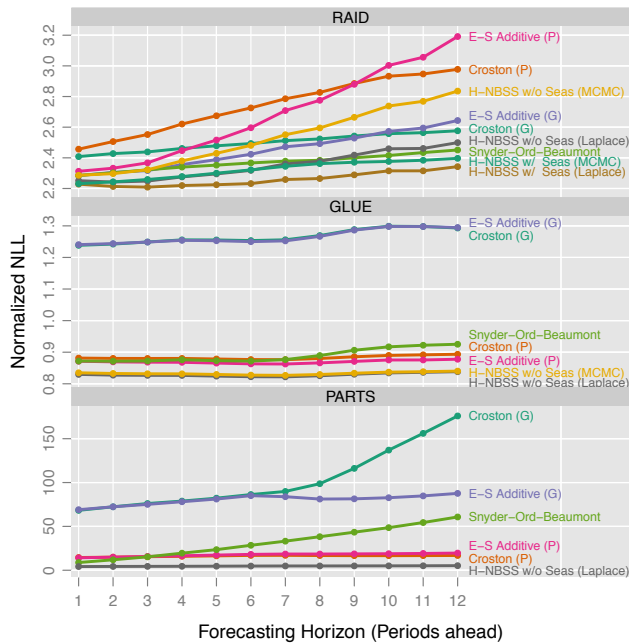
Figure 5. Average out-of-sample NLL for all datasets as a function of the forecasting horizon. The proposed H-NBSS model with the Laplace approximation exhibits the best performance.

Table 3. Forecasting performance for the masked series in the RAID dataset, for which only four observations are available.

	Normalized NLL	Relative MSE	Relative MAE
Independent models	79.04	7.69	2.13
Hierarchical model	**72.77**	**4.13**	**1.75**

measures predictive distributional accuracy) the H-NBSS model consistently yields the best performance, with the Laplace approximation slightly beating MCMC (both approximations are very close). The MSE and MAE measures tell a consistent story, the only exception being the PARTS dataset where Croston very slightly bests the other approaches in the forecast of the mean (MSE measure). This can be explained by high proportion of "lumpy" series in PARTS (*cf.* Table 1), which exhibit high demand variability, and hence low predictability.

4.3. Benefits of Hierarchy

We close this section by outlining the benefits of the hierarchical structure in H-NBSS. The major advantage of sharing information across several series in a group lies in the ability to increase "statistical strength", in particular for series for which little history is available. We illustrate this
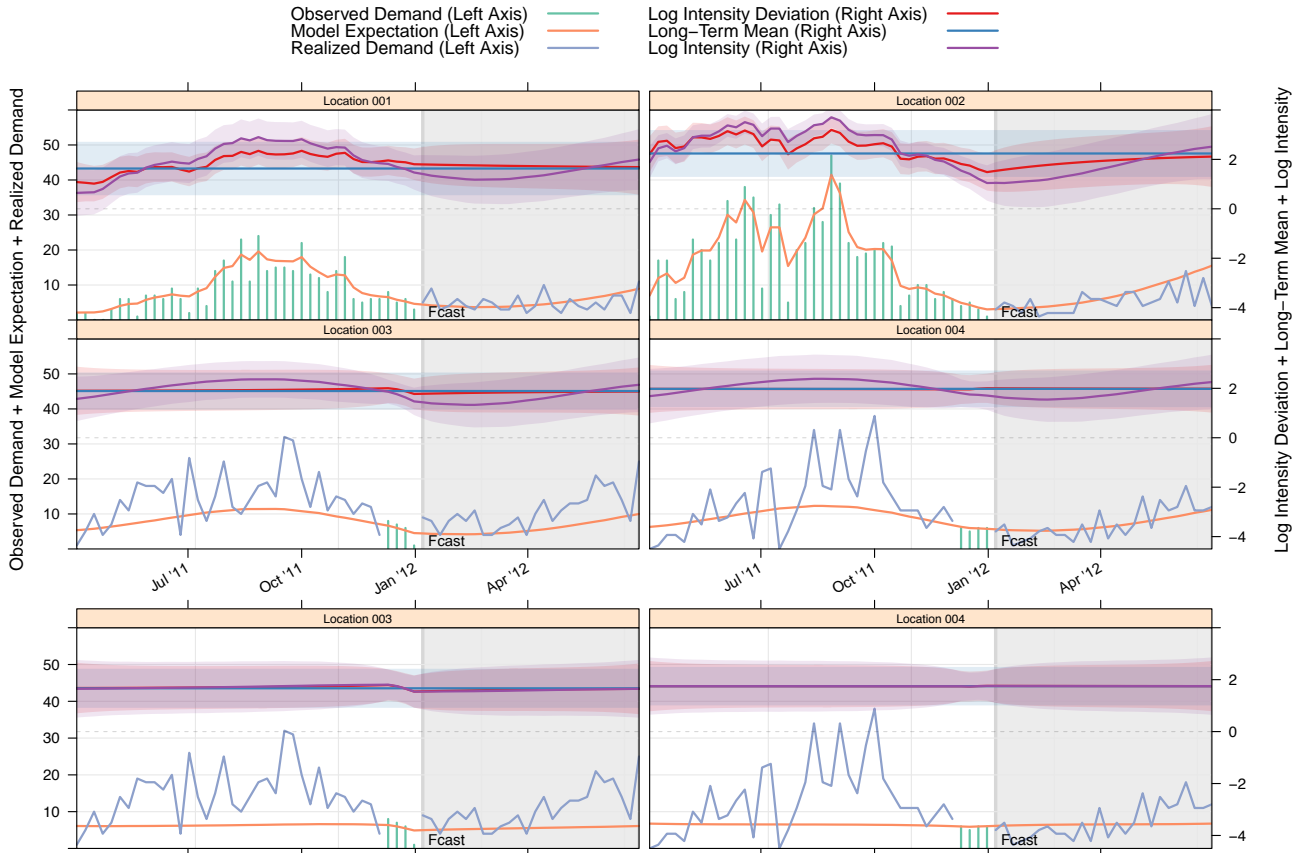
Figure 6. **Top 4 panels:** Example of information sharing in the hierarchical model. The time series for 4 stores of the RAID dataset are included in the group, the first two with a full history (green vertical lines) and the bottom two with only four observations of history each (which appear just before the start of the forecasting period). The realized observations are indicated by the solid blue lines and the expectation under the model distribution is the orange line; we note that the model deduces sensible seasonalities over the forecasting horizon for Locations 3 and 4 even with only four observations in their respective histories, and can reasonably "backcast" over the missing history as well. **Bottom 2 panels:** Results with an independent model for each series (no sharing); with only four observations in the history, the models cannot do much better than fit a constant.

in Fig. 6, which shows that when groups of time series can share information, useful patterns can be learned even for series with very short histories (here, four observations). In contrast, with no sharing, the model cannot do much better than fit a constant.

These results translate quantitatively in Table 3. On 20 of the 24 series of the RAID dataset, we provided only the four observations before the forecasting horizon; for the other 4 series, we provided the full history. The table contrasts the performance of a H-NBSS model trained separately for each series ("independent models"), versus a single H-NBSS model grouping the 24 series together ("hierarchical model") and reports average performance only on the 20 masked series. There is a dramatic gain in performance attributable to sharing in the hierarchical model: seasonalities learned on the four complete series transfer to the incomplete ones.

5. Conclusion

This paper introduced a modeling methodology for groups of related time series of counts, such as the small-integer series frequently encountered in supply chain operations. We outlined the sizable accuracy gains possible through jointly modeling several time series in a hierarchical Bayesian framework and presented an effective approximate inference algorithm to make the H-NBSS model useful in practice. Future work should investigate other trade-offs on the accuracy–computational cost spectrum, such as the INLA approach (Rue et al., 2009), recently revisited by Han et al. (2013). Beyond its increased accuracy, the H-NBSS model can provide real-world benefits in applications where count data dominate, for instance by supplying useful forecasts for new stores with very little (or no) history and improving the efficiency of inventory management policies with better distributions of future demand.

References

Altay, N. and Litteral, L. A. (eds.). *Service Parts Management: Demand Forecasting and Inventory Control*. Springer, 2011.

Barber, D., Cemgil, A. T., and Chiappa, S. (eds.). *Bayesian time series models*. Cambridge University Press, 2011.

Bishop, C. *Pattern recognition and machine learning*. Springer, New York, 2006.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. *Time series analysis: forecasting and control*. John Wiley & Sons, fourth edition, 2008.

Croston, J. D. Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3):289–303, September 1972.

Davis, T. A. *Direct methods for sparse linear systems*. SIAM, 2006.

Deisenroth, M. and Mohamed, S. Expectation propagation in Gaussian process dynamical systems. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2618–2626. 2012.

Doucet, A., De Freitas, N., and Gordon, N. (eds.). *Sequential Monte Carlo methods in practice*. Springer New York, 2001.

Durbin, J. and Koopman, S. J. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society: Series B*, 62(1):3–56, 2000.

Durbin, J. and Koopman, S. J. *Time series analysis by state space methods*. Oxford University Press, second edition, 2012.

Fox, E., Sudderth, E., Jordan, M. I., and Willsky, A. S. Sharing features among dynamical systems with beta processes. In Y. Bengio, D. Schuurmans, J. Lafferty and Williams, C. (eds.), *Advances in Neural Information Processing Systems 22*, 2010.

Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

Han, S., Liao, X., and Carin, L. Integrated non-factorized variational inference. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2481–2489. 2013.

Heskes, T. and Zoeter, O. Expectation propagation for approximate inference in dynamic bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 216–223. Morgan Kaufmann Publishers Inc., 2002.

Hilbe, J. M. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK, second edition, 2011.

Hoffman, M. D. and Gelman, A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, in press, 2013.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. *Forecasting with exponential smoothing: the state space approach*. Springer, Berlin and Heidelberg, 2008.

Hyndman, R. J., Athanasopoulos, G., Razbash, S., Schmidt, D., Zhou, Z., and Khan, Y. *forecast: Forecasting functions for time series and linear models*, 2013. URL http://CRAN.R-project.org/package=forecast. R package version 4.04.

Kitagawa, G. Non-Gaussian statespace modeling of nonstationary time series. *Journal of the American statistical association*, 82 (400):1032–1041, 1987.

Kwan, H. W. *On the Demand Distributions of Slow-Moving Items*. PhD thesis, University of Lancaster, UK, 1991.

Lambert, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Maybeck, P. S. *Stochastic Models, Estimation, and Control, Volume 1*. Mathematics in Science and Engineering, Volume 141. Academic Press, 1979.

Minka, T. P. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, 1999.

Robert, C. P., Ryden, T., and Titterington, D. M. Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *Journal of Statistical Computation and Simulation*, 64:327–355, 1999.

Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics and Applied Probability 104. Chapman & Hall/CRC, 2005.

Rue, H., Martino, S., and Chopin, N. Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Scott, S. L. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

Shenstone, L. and Hyndman, R. J. Stochastic models underlying Croston's method for intermittent demand forecasting. *Journal of Forecasting*, 24(6):389–402, 2005.

Snyder, R. D., Ord, J. K., and Beaumont, A. Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2):485–496, 2012.

Stan Development Team. Stan: A C++ library for probability and sampling, 2013. URL http://mc-stan.org/.

Syntetos, A. A., Boylan, J. E., and Croston, J. D. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5):495–503, 2005.

Syntetos, A A, Babai, M Z, and Altay, N. On the demand distributions of spare parts. *International Journal of Production Research*, 50(8):2101–2117, April 2012.

Teh, Y. W. and Jordan, M. I. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Mueller and Walker, S. (eds.), *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

Yu, B M, Shenoy, K V, and Sahani, M. Expectation propagation for inference in non-linear dynamical models with poisson observations. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, pp. 83–86. IEEE, 2006.