Latent Gaussian Count Time Series

Yisu Jia                             Stefanos Kechagias
University of North Florida         SAS Institute
James Livsey                        Robert Lund *
United States Census Bureau         University of California - Santa Cruz
                                      Vladas Pipiras †
                                      University of North Carolina - Chapel Hill

June 7, 2021

Abstract

This paper develops the theory and methods for modeling a stationary count time series via Gaussian transformations. The techniques use a latent Gaussian process and a distributional transformation to construct stationary series with very flexible correlation features that can have any pre-specified marginal distribution, including the classical Poisson, generalized Poisson, negative binomial, and binomial structures. Gaussian pseudo-likelihood and implied Yule-Walker estimation paradigms, based on the autocovariance function of the count series, are developed via a new Hermite expansion. Particle filtering and sequential Monte Carlo methods are used to conduct likelihood estimation. Connections to state space models are made. Our estimation approaches are evaluated in a simulation study and the methods are used to analyze a count series of weekly retail sales.

Keywords: Count Distributions; Hermite Expansions; Likelihood Estimation; Particle Filtering; Sequential Monte Carlo; State Space Models

*Robert Lund’s research was partially supported by the grant NSF DMS 1407480.
†Vladas Pipiras’s research was partially supported by the grant NSF DMS 1712966.
1 Introduction

This paper develops the theory and methods for modeling a stationary discrete-valued time series by transforming a Gaussian process. Since the majority of discrete-valued time series involve integer counts supported on some subset of \{0, 1, \ldots\}, we isolate on this support set. Our methods are based on a copula-style transformation of a latent Gaussian stationary series and are able to produce any desired count marginal distribution. It is shown that the proposed model class produces the most flexible pairwise correlation structures possible, including negatively dependent series. Model parameters are estimated via 1) a Gaussian pseudo-likelihood approach, developed from some new Hermite expansion techniques, which use only the mean and the autocovariance of the series, 2) an implied Yule-Walker moment estimation approach when the latent Gaussian process is an autoregression, and 3) a particle filtering (PF) / sequential Monte Carlo (SMC) approach that uses a state space model (SSM) representation of the transformation to approximate the true likelihood. Extensions to non-stationary settings, particularly those with covariates, are discussed.

The theory of stationary Gaussian time series is by now well developed. A central result is that a stationary Gaussian series \{X_t\}_{t \in \mathbb{Z}} having the lag-\(h\) autocovariance \(\gamma_X(h) = \text{Cov}(X_t, X_{t+h})\) exist if and only if \(\gamma_X\) is symmetric about lag zero and non-negative definite (see Theorem 1.5.1 in [6]). However, such a result does not hold for stationary count series having a certain prescribed marginal distribution (e.g, Poisson). In principle, distributional existence issues are checked with Kolmogorov’s consistency criterion (see Theorem 1.2.1 in [6]); in practice, one needs a specified joint distribution to check for consistency. Phrased another way, Kolmogorov’s consistency criterion is not a constructive result and does not illuminate how to build stationary time series having a particular marginal distribution and correlation structure. Perhaps owing to this, count time series have been constructed
from a plethora of approaches over the years, as is next reviewed.

Drawing from the success of autoregressive moving-average (ARMA) models in describing stationary Gaussian series, early count authors constructed correlated count series from discrete ARMA (DARMA) and integer ARMA (INARMA) difference equation methods. Focusing on the first order autoregressive case for simplicity, a DAR(1) series \( \{ X_t \} \) with specified marginal distribution \( F_X(\cdot) \) is obtained by generating \( X_1 \) from \( F_X(\cdot) \) and then at each subsequent time, either keeping the previous count value with probability \( p \) or generating an independent copy of \( F_X(\cdot) \) with probability \( 1 - p \). INAR(1) series are built via the thinned AR(1) equation \( X_t = p \circ X_{t-1} + \epsilon_t \), where \( \{ \epsilon_t \} \) is an IID count-valued random sequence and \( \circ \) is a thinning operator defined by \( p \circ Y = B(Y, p) \) for a binomial distribution \( B(n, p) \) with \( n \) trials and success probability \( p \). DARMA methods were initially explored in [23], but were subsequently discarded by practitioners because their sample paths often remained constant for long periods, especially in highly correlated cases; INARMA series are still used today. In contrast to their Gaussian ARMA brethren, DARMA and INARMA models, and their extensions in [24], cannot produce negative autocorrelations.

The works [5] and [9] take a different approach, producing the desired count marginal distribution by combining IID copies of a correlated Bernoulli series \( \{ B_t \} \) built from a stationary renewal sequence. Explicit autocovariance functions when \( \{ B_t \} \) is made by binning (clipping) a stationary Gaussian sequence into zero-one categories are derived in [33]. While these models can have negative correlations, they do not necessarily produce the most negatively correlated count structures possible. Also, some important count marginal distributions, including generalized Poisson, are not easily built from these methods. The results here easily generate any desired count marginal distribution. Other count model classes studied include Gaussian processes rounded to their nearest integer [26], hierarchical
Bayesian count model approaches [2], and others (see [18] and [11] for recent reviews). Each approach has some drawbacks.

The models here impose a fixed marginal distribution for the counts. This is in contrast to generalized ARMA methods (GLARMA), which typically posit conditional distributions in lieu of marginal distributions, with model parameters typically being random. As [1] shows in the Poisson case, once the randomness of the parameters is taken into account, the true marginal distribution of the series can be far from the posited conditional distribution. This said, the literature on GLARMA and other conditional models is extensive [3, 43]. See [16] for a recent review of GLARMA models.

A time series analyst generally needs four features in a count model: 1) general marginal distributions; 2) the most general correlation structures possible, both positive and negative; 3) the straight-forward accommodation of covariates; and 4) a well performing and computationally feasible likelihood inference approach. All previous count classes fail to accommodate one or more of these tenets. This paper’s purpose is to introduce and study a count model class that, for the first time, simultaneously achieves all four features. Our model employs a latent Gaussian process and a copula-style transformation. This type of construction has recently shown promise in spatial statistics [12, 21], multivariate modeling [39, 40], and regression [35], but the theory has yet to be developed for count series ([35, 30 provide some partial results). Our objectives here are several-fold. On a methodological level, it is shown, through some newly derived Hermite polynomial expansions, that accurate and efficient numerical quantification of the correlation structure of this count model class is feasible. Based on a result in [42], the class is shown to produce the most flexible pairwise correlation structures possible, positive or negative (see Remark 2.2 below). Connections to both importance sampling schemes, where the popular GHK sampler in [35] is
adapted to our needs, and to the SSM and SMC literature, which allow natural extensions of the GHK sampler and likelihood evaluation, are made. The methods are tested on both synthetic and real data.

The works [35, 30] are perhaps the closest papers to this study. While the general latent Gaussian construct adopted is the same, our work differs in that explicit autocovariance relations are developed via Hermite expansions, flexibility and optimality issues of the model class are addressed, Gaussian pseudo-likelihood and implied least-squares parameter estimation approaches are developed, and both the importance sampling and SSM connections are explored in detail. Additional connections to [35, 30] and to the spatial count modeling papers [21, 22] are later made.

The rest of this paper proceeds as follows. The next section and Appendix A introduce our Gaussian transformation count model and establish its basic mathematical and statistical properties. Section 3 and Appendix B move to estimation, developing three techniques: a Gaussian pseudo-likelihood approach, implied Yule-Walker estimation, and PF/SMC methods. Section 4 and Appendix C present simulation results. Section 5 and Appendix D analyze soft drink sales counts at one location of the now defunct Dominicks Finer Foods retail chain. This series exhibits overdispersion, negative lag one autocorrelation, and dependence on a price reduction (sales) covariate, which illustrates the flexibility of our approach. Section 6 concludes with comments and suggestions for future research.

2 Theory

We seek to construct a strictly stationary time series \( \{X_t\} \) having marginal distributions from any family of count distributions supported in \( \{0, 1, \ldots\} \), including the Binomial,
Poisson, mixture Poisson, negative binomial, generalized Poisson, and Conway-Maxwell-Poisson distributions. The later three distributions are over-dispersed (their variances are larger than their respective means), which is the case for many observed count time series.

Let \( \{X_t\}_{t \in \mathbb{Z}} \) be the stationary count time series of interest. Suppose that one wants the marginal cumulative distribution function (CDF) of \( X_t \) for each \( t \) of interest to be \( F_X(x) = \mathbb{P}[X_t \leq x] \), depending on a vector \( \theta \) containing all CDF model parameters. The series \( \{X_t\} \) will be modeled through

\[
X_t = G(Z_t), \quad \text{where} \quad G(z) = F_X^{-1}(\Phi(z)), \quad z \in \mathbb{R},
\]

and \( \Phi(\cdot) \) is the CDF of a standard normal variable and \( F_X^{-1}(u) = \inf\{t : F_X(t) \geq u\}, \quad u \in (0,1) \), is the generalized inverse (quantile function) of the CDF \( F_X \). The process \( \{Z_t\}_{t \in \mathbb{Z}} \) is standard Gaussian for each fixed \( t \), but possibly correlated in time:

\[
\mathbb{E}[Z_t] = 0, \quad \mathbb{E}[Z_t^2] = 1, \quad \rho_Z(h) =: \text{Corr}(Z_t, Z_{t+h}) = \mathbb{E}[Z_t Z_{t+h}].
\]

This approach has been used in [39, 35, 21, 30] with good results. The autocovariance function (ACVF) of \( \{Z_t\} \), denoted by \( \gamma_Z(\cdot) \), is the same as the autocorrelation function (ACF) due to standard normality and depends on another vector \( \eta \) of ACVF parameters.

As expanded on in Section 2.3, (1) can be viewed as a SSM:

State equation : \( p(z_t|z_{t-1}, \ldots, z_1) \) governing latent Gaussian dynamics;

Observation equation : \( \mathbb{P}(X_t = k|z_t) = 1_{A_k}(z_t) \) with the set \( A_k \) defined below.

Here, \( p(\cdot|\cdot) \) is notation for an arbitrary conditional distribution.

This model has alternative names in other literature. For example, [7] call this setup the normal to anything (NORTA) procedure in operations research, whereas [20] calls this
a translational model in mechanical engineering. Our goal is to give a reasonably complete analysis of the probabilistic and statistical properties of these models.

The construction in (1) ensures that the marginal CDF of $X_t$ is indeed $F_X(\cdot)$. Elaborating, the probability integral transformation theorem shows that $\Phi(Z_t)$ has a uniform distribution over $(0,1)$ for each $t$; a second application of the result justifies that $X_t$ has the marginal distribution $F_X(\cdot)$ for each $t$. Moreover, temporal dependence in $\{Z_t\}$ will induce temporal dependence in $\{X_t\}$ as quantified below. For notation, let $\gamma_X(h) = E[X_{t+h}X_t] - E[X_{t+h}]E[X_t]$ denote the ACVF of $\{X_t\}$.

2.1 Relationship between autocovariances

The autocovariance functions of $\{X_t\}$ and $\{Z_t\}$ can be related using Hermite expansions (see Chapter 5 of [37]). In particular, using the Hermite polynomials $H_k(z) = (-1)^k e^{z^2/2} \frac{d^k}{dz^k} (e^{-z^2/2})$, $z \in \mathbb{R}$ we can expand the $L^2$ function $G$ as

$$G(z) = E[G(Z_0)] + \sum_{k=1}^{\infty} g_k H_k(z)$$

(3)

where the Hermite coefficients $g_k$ are given by

$$g_k = \frac{1}{k!} \int_{-\infty}^{\infty} G(z)H_k(z) \frac{e^{-z^2/2}dz}{\sqrt{2\pi}} = \frac{1}{k!} E[G(Z_0)H_k(Z_0)],$$

(4)

for a standard normal variable $Z_0$. The relationship between $\gamma_X(\cdot)$ and $\gamma_Z(\cdot)$ is key and is extracted from Chapter 5 of [37]:

$$\gamma_X(h) = \sum_{k=1}^{\infty} k! g_k^2 \gamma_Z(h)^k =: g(\gamma_Z(h)),$$

(5)
where \( g(u) = \sum_{k=1}^{\infty} k! g_k^2 u^k \). For \( h = 0 \), (5) yields \( \text{Var}(X_t) = \gamma_X(0) = \sum_{k=1}^{\infty} k! g_k^2 \), which depends only on the marginal parameters in \( \theta \). Moreover, the ACF of \( \{X_t\} \) is

\[
\rho_X(h) = \sum_{k=1}^{\infty} \frac{k! g_k^2}{\gamma_X(0)} \gamma_Z(h)^k =: L(\rho_Z(h)),
\]

where

\[
L(u) = \sum_{k=1}^{\infty} \frac{k! g_k^2}{\gamma_X(0)} u^k =: \sum_{k=1}^{\infty} \ell_k u^k,
\]

and \( \ell_k = k! g_k^2 / \gamma_X(0) \). The function \( L \) maps \([-1, 1]\) into (but not necessarily onto) \([-1, 1]\).

For future reference, note that \( L(0) = 0 \) and \( L(1) = \sum_{k=1}^{\infty} \ell_k = 1 \). Using (3) and \( \mathbb{E}[H_k(Z_0)H_l(-Z_0)] = (-1)^k k! [k;1] \) gives \( L(-1) = \text{Corr}(G(Z_0), G(-Z_0)) \); however, \( L(-1) \) is not necessarily \(-1\) in general. As such, \( L(\cdot) \) “starts” at \((-1, L(-1))\), passes through \((0, 0)\), and connects to \((1, 1)\). Examples are given in Figure 2 of Appendix A.

We call the quantity \( L(\cdot) \) a link function, and the coefficients \( \ell_k, k \geq 1 \), link coefficients. (Sometimes, slightly abusing terminology, we also use these terms for \( g(\cdot) \) and \( g_k^2 k! \), respectively.) A key feature in (5) is that the effects of the marginal CDF \( F_X(\cdot) \) and the ACVF \( \gamma_Z(\cdot) \) are “decoupled” in the sense that the correlation parameters in \( \{Z_t\} \) do not influence the \( g_k \) coefficients in (5) — this is useful later in estimation.

Further properties and the numerical calculation of the link function and the Hermite coefficients are discussed in Appendix A. The computation of the Hermite coefficients, in particular, is feasible due to the following lemma, which is proved in Appendix A.

**Lemma 2.1.** If \( \mathbb{E}[X_t^p] < \infty \) for some \( p > 1 \), then the coefficients \( g_k \) satisfy

\[
g_k = \frac{1}{k! \sqrt{2\pi}} \sum_{n=0}^{\infty} e^{-\Phi^{-1}(C_n)^2/2} H_{k-1}(\Phi^{-1}(C_n)),
\]

where \( C_n = \mathbb{P}[X_t \leq n] \). (When \( \Phi^{-1}(C_n) = \pm \infty \) (that is, \( C_n = 0 \) or \( 1 \)), the summand \( e^{-\Phi^{-1}(C_n)^2/2} H_{k-1}(\Phi^{-1}(C_n)) \) is interpreted as zero.)
Returning to the relationship between $\rho_X(h)$ and $\rho_Z(h)$, from (6), one can see that

\[ |\rho_X(h)| \leq |\rho_Z(h)|, \tag{9} \]

which implies that a positive $\rho_Z(h)$ leads to a positive $\rho_X(h)$. A negative $\rho_Z(h)$ produces a negative $\rho_X(h)$ since $L(u)$ is, in fact, monotone increasing (see Proposition A.1 in Appendix A) and crosses zero at $u = 0$ (the negativeness of $\rho_X(h)$ when $\rho_Z(h) < 0$ can also be deduced from the nondecreasing nature of $G$ via an inequality on page 20 of [41] for Gaussian variables).

**Remark 2.1.** The short- and long-range dependence properties of $\{X_t\}$ can be extracted from those of $\{Z_t\}$. Recall that a time series $\{Z_t\}$ is short-range dependent (SRD) if $\sum_{h=-\infty}^{\infty} |\rho_Z(h)| < \infty$. According to one definition, a series $\{Z_t\}$ is long-range dependent (LRD) if $\rho_Z(h) = Q(h)h^{2d-1}$, where $d \in (0, 1/2)$ is the LRD parameter and $Q$ is a slowly varying function at infinity [37]. The ACVF of such LRD series satisfies $\sum_{h=-\infty}^{\infty} |\rho_Z(h)| = \infty$. If $\{Z_t\}$ is SRD, then so is $\{X_t\}$ by (9). On the other hand, if $\{Z_t\}$ is LRD with parameter $d$, then $\{X_t\}$ can be either LRD or SRD. The conclusion depends, in part, on the Hermite rank of $G(\cdot)$, which is defined as $r = \min\{k \geq 1 : g_k \neq 0\}$. Specifically, if $d \in (0, (r - 1)/2r)$, then $\{X_t\}$ is SRD; if $d \in ((r - 1)/2r, 1/2)$, then $\{X_t\}$ is LRD with parameter $r(1/2) + 1/2$ (see [37], Proposition 5.2.4).

The model in (1) admits the following structure: if $Z_t$ and $Z_s$ are independent, then so are $X_t$ and $X_s$. It follows that if $\{Z_t\}$ is stationary and $q$-dependent, than both $\{Z_t\}$ and $\{X_t\}$ must be $q$th order moving-average time series. Unfortunately, no analogous autoregressive structure holds; in fact, if $\{Z_t\}$ is a first order autoregression, then $\{X_t\}$ may not be an autoregression of any order (this can be inferred from [28]).
Remark 2.2. The construction in (1) yields models with the most flexible correlations possible for \( \text{Corr}(X_{t_1}, X_{t_2}) \) for two variables \( X_{t_1} \) and \( X_{t_2} \) with the same marginal distribution \( F_X \). Indeed, let \( \rho_− = \min\{\text{Corr}(X_{t_1}, X_{t_2}) : X_{t_1}, X_{t_2} \sim F_X\} \) and define \( \rho_+ \) similarly with min replaced by max. Then, as shown in Theorem 2.5 of [42],

\[
\rho_+ = \text{Corr}(F_X^{-1}(U), F_X^{-1}(U)) = 1, \quad \rho_- = \text{Corr}(F_X^{-1}(U), F_X^{-1}(1-U)),
\]

where \( U \) is a uniform random variable over \((0,1)\). Since \( U \overset{D}= \Phi(Z) \) and \( 1−U \overset{D}= \Phi(−Z) \) for a standard normal random variable \( Z \), the maximum and minimum correlations \( \rho_+ \) and \( \rho_- \) are indeed achieved with (1) when \( Z_{t_1} = Z_{t_2} \) and \( Z_{t_1} = −Z_{t_2} \), respectively. The preceding statements are non-trivial for \( \rho_- \) only since \( \rho_+ = 1 \) is attained whenever \( X_{t_1} = X_{t_2} \). It is worthwhile to compare this to the discussion following (7). Finally, all correlations in \((\rho_−, \rho_+) = (\rho_−, 1)\) are achievable since \( L(u) \) in (7) is continuous in \( u \). The flexibility of correlations for Gaussian copula models in the spatial context was also noted and studied in [21], especially in comparison to a class of hierarchical, e.g. Poisson, models.

The preceding remark settles autocovariance flexibility issues for stationary count series. Flexibility is a concern when the series is negatively correlated, an issue arising, for example, with hurricane counts in [33] and chemical process counts in [26]. Since any general count marginal distribution can also be achieved, the model class is quite general.

### 2.2 Covariates

There are situations where stationarity is not desired. Such scenarios can often be accommodated by simple variants of the above setup. For concreteness, consider a situation where a vector \( M_t \) of \( J \) non-random covariates is available to explain the series at time \( t \). If one wants \( X_t \) to have the marginal distribution \( F_{\theta(t)}(\cdot) \), where \( \theta(t) \) is a vector-valued
function of $t$ containing marginal distribution parameters, then simply set

$$X_t = F_{\theta(t)}^{-1}(\Phi(Z_t))$$

(10)

and reason as before. We do not recommend modifying $\{Z_t\}$ for the covariates as this may bring process existence issues into play.

Generalized linear models link functions (not to be confused with $L(\cdot)$ in (6)–(7)) can be used when parametric support set bounds are encountered. For example, a Poisson regression with correlated errors can be formulated via a parameter vector $\beta$ of regression coefficients with $\theta(t) = \mathbb{E}[X_t] = \exp(\beta' M_t)$. Here, the exponential link guarantees that the Poisson parameter is positive. The above construct requires the covariates to be non-random; should covariates be random, marginal distributions may change from $F_{\theta(t)}$.

2.3 Particle filtering and state space model connections

This subsection studies the implications of the latent structure of our model, especially as it relates to SSMs and importance sampling approaches. This will be used to construct PF/SMC approximations of various quantities, and in goodness-of-fit assessments. Our main reference is [14]. As in that monograph, let $z_{0:t} = \{Z_0 = z_0, \ldots, Z_t = z_t\}$, $x_{0:t} = \{X_0 = x_0, \ldots, X_t = x_t\}$, and $p(\cdot)$ and $p(\cdot|\cdot)$ denote joint and conditional probabilities (or their densities, depending on the context). For example, $p(z_{0:t}|x_{0:t})$ denotes the conditional density of $Z_{0:t}$ given $x_{0:t}$. Similarly, let $\mathbb{E}[\cdot|x_{0:t}]$ denote conditional expectation given $x_{0:t}$.

The SSM formulation starts by specifying $p(z_{t+1}|z_{0:t})$ and $p(x_t|z_t)$. While $\{Z_t\}$ is often first order Markov, implying that $p(z_{t+1}|z_{0:t}) = p(z_{t+1}|z_t)$, this is not necessary.

To specify $p(z_{t+1}|z_{0:t})$ in our stationary Gaussian case, we compute the best one-step-ahead linear prediction of $Z_{t+1}$ from $z_{0:t}$ given by $\hat{Z}_{t+1} = \phi_{t0} Z_t + \ldots + \phi_{tt} Z_0$. The coefficients
\( \phi_{ts}, s \in \{0, \ldots, t\}, \) can be computed recursively in \( t \) from the ACF of \( \{Z_t\} \) via the classical Durbin-Levinson (DL) or the Innovations algorithm, for examples. As a convention, we take \( \hat{Z}_0 = 0. \) Let \( r_t^2 = \mathbb{E}[(Z_t - \hat{Z}_t)^2] \) be the corresponding unconditional mean squared prediction error. With this notation,

\[
p(\hat{z}_{t+1}|z_{0:t}) \overset{D}{=} \mathcal{N}(\hat{z}_{t+1}, r_{t+1}^2),
\]

where \( \hat{z}_{t+1} = \phi_{t0}z_t + \ldots + \phi_{tt}z_0. \) Again, \( \{Z_t\} \) does not have to be Markovian (of any order).

On the other hand, with (1),

\[
p(x_t|z_t) = \delta_{G(z_t)}(x_t) = \begin{cases} 1, & \text{if } x_t = G(z_t), \\ 0, & \text{otherwise}, \end{cases}
\]

where \( \delta_y(x) \) is a unit point mass at \( \{y\}. \) The equations in (11) and (12) constitute the SSM representation of (1).

In inference and related tasks for SSMs, the basic goal is to compute the conditional expectation \( \mathbb{E}[v(Z_{0:t})|x_{0:t}] \) for some function \( v. \) This is often carried out through an importance sampling algorithm such as sequential importance sampling (SIS), which generates \( N \) independent particle trajectories \( Z^i_{0:t}, i \in \{1, \ldots, N\}, \) from a proposal distribution \( \pi(z_{0:t}|x_{0:t}) \) and approximates the conditional expectation as

\[
\mathbb{E}[v(Z_{0:t})|x_{0:t}] \approx \sum_{i=1}^{N} v(Z^i_{0:t}) \tilde{w}_i =: \hat{\mathbb{E}}[v(Z_{0:t})|x_{0:t}],
\]

where

\[
\tilde{w}_i = \frac{w(Z^i_{0:t})}{\sum_{i=1}^{N} w(Z^i_{0:t})}, \quad w(z_{0:t}) = \frac{p(z_{0:t}|x_{0:t})}{\pi(z_{0:t}|x_{0:t})},
\]

are the (normalized) importance weights (see [14] and [32]). Furthermore, in SIS,

\[
\tilde{w}_i \propto \tilde{w}_{i-1} w_t(Z^i_{0:t}), \quad w_t(z_{0:t}) = \frac{p(x_t|z_t)p(z_t|z_{0:t-1})}{\pi(z_t|z_{0:t-1}, x_{0:t})}
\]

12
(see (1.6) in [14], which is adapted to a possibly non-Markov setting by replacing \( p(z_t|z_{t-1}) \) with \( p(z_t|z_{0:t-1}) \)). The two probability terms in the numerator of \( w_t(z_{0:t}) \) in (15) constitute the SSM, whereas the denominator relates to the proposal distribution.

We suggest the following proposal distribution and the resulting SIS algorithm for our model. Take

\[
\pi(z_t|x_{0:t}) \triangleq \mathcal{N}_{A_{x_t}}(\hat{z}_t, r_t^2),
\]

where \( \mathcal{N}_{A} \) denotes a normal distribution restricted to the set \( A \), and

\[
A_k = \{ z : \Phi^{-1}(C_{k-1}) \leq z \leq \Phi^{-1}(C_k) \}.
\]

The role of \( A_k \) stems from the fact

\[
k = G(z) \iff z \in A_k
\]

(i.e., the count value \( k \) is obtained if and only if \( Z_t \in A_k \); see the expression (A.2) for \( G(z) \)). In particular, for \( \{Z_t^i\} \) generated from the proposal distribution (16), the term \( p(x_t|Z_t^i) \) in the incremental weight \( w_t(Z_{0:t}^i) \) of (15) is always set to unity. The rest of the incremental weights are calculated as

\[
w_t(z_{0:t}) = \frac{p(z_t|z_{0:t-1})}{\pi(z_t|z_{0:t-1}, x_{0:t})} = \frac{e^{-\frac{(z_t-\hat{z}_t)^2}{2r_t^2}}/(2\pi r_t^2)^{1/2}}{e^{-\frac{(z_t-\hat{z}_t)^2}{2r_t^2}}/[2(2\pi r_t^2)^{1/2} \times \mathbb{P}(N(\hat{\mu}_t, r_t^2) \in A_{x_t})]} = \mathbb{P}(N(\hat{\mu}_t, r_t^2) \in A_{x_t}) = \Phi\left(\frac{\Phi^{-1}(C_{x_t}) - \hat{\mu}_t}{r_t}\right) - \Phi\left(\frac{\Phi^{-1}(C_{x_{t-1}}) - \hat{\mu}_t}{r_t}\right) =: w_t(\hat{\mu}_t).
\]

The choice of the proposal distribution is largely motivated by \( \mathbb{P}(X_t = k|Z_t^i) = 1_{A_k}(Z_t^i) \) and the explicit form in (19) for the incremental weights \( w_t(Z_{0:t}) \). Optimality considerations are mentioned in Remark B.3.

The following steps summarize our SIS algorithm.
**Sequential Importance Sampling (SIS):** For \( i \in \{1, \ldots, N\} \), where \( N \) represents the number of particles, initialize the weight \( w_i^0 = 1 \) and the latent series \( Z_0^i \) by

\[
Z_0^i \sim \mathcal{N}_{A_{x_0}}(0, 1).
\]  

(20)

Then, recursively over \( t = 1, \ldots, T \), perform the following steps:

1: Compute \( \hat{Z}_t^i \) with the DL or other algorithm using the previously generated values of \( Z_0^i, \ldots, Z_{t-1}^i \).

2: Update the series \( Z_t^i \) and the importance weight \( w_t^i \) via

\[
Z_t^i \sim \mathcal{N}_{A_{x_t}}(\hat{Z}_t^i, r_t^2), \quad w_t^i = w_{t-1}^i w_t(\hat{Z}_t^i),
\]

(21)

where \( w_t(z) \) is defined in (19).

**Remark 2.3.** For \( i \in \{1, \ldots, N\} \), the constructed path \( \{Z_t^i\}_{t=0}^T \) is one of the \( N \) independent “particles” used to approximate the conditional expectation in (13). Equation (21) ensures that for each \( i \), the path \( \{Z_t^i\}_{t=0}^T \) obeys the restriction \( G(Z_t^i) = x_t \) and matches the temporal structure of \( \{Z_t\} \). These two properties show that \( \{Z_t^i\}_{t=0}^T \) is a realization of the latent Gaussian stationary series producing \( X_t = x_t \) for all \( t \). Finally, we note where the model parameters enter into the SIS algorithm. The marginal distribution parameters \( \theta \) enter through the form of \( C_x \) in (19), whereas the temporal dependence parameters \( \eta \) enter through the one-step-ahead prediction coefficients \( \phi_{ts}, s \in \{0, \ldots, t\} \), in the calculation of \( \hat{Z}_t^i \) in Step 1 of the algorithm, and through the prediction error \( r_t \).

To compute the model likelihood, several known formulas applicable in the (general) SIS setting are needed. The relation

\[
\frac{p(z_{0:t}|x_{0:t})}{\pi(z_{0:t}|x_{0:t})} = \prod_{s=0}^t w_s(z_{0:s}) \frac{p(x_0)}{p(x_{0:t})},
\]

14
produces
\[ E[w_i^t v(Z_{0:t}^i)|x_{0:t}] p(x_0) = E[v(Z_{0:t})|x_{0:t}] p(x_0). \]

In particular (with \( v(\cdot) \equiv 1 \)),
\[ E[w_i^t] p(x_0) = p(x_{0:t}). \quad (22) \]

To conduct prediction, we use Equation (1.2) in [14] to get
\[ E[v(X_{t+1})|x_{0:t}] = E\left[ E[v(G(\hat{Z}_{t+1}))|Z_{0:t}]|x_{0:t} \right] =: E[D_{v,t+1}(\hat{Z}_{t+1})|x_{0:t}], \quad (23) \]

where
\[ D_{v,t+1}(z) = E\left[v(G(N(z,r_{t+1}^2)))\right] = \int_{\mathbb{R}} v(G(z_{t+1})) \frac{1}{\sqrt{2\pi r_{t+1}^2}} e^{-\frac{(z_{t+1}-z)^2}{2r_{t+1}^2}} dz_{t+1}, \quad (24) \]
since \( \hat{Z}_{t+1}|z_{0:t} \overset{D}{=} N(\hat{z}_{t+1}, r_{t+1}^2) \). In view of (23) and (13), the following prediction approximation arises:
\[ E[v(X_{t+1})|x_{0:t}] \approx \sum_{i=1}^{N} \frac{w_i^t}{\Omega_{N,t}} D_{v,t+1}(\hat{Z}_{t+1}^i) =: \hat{E}[v(X_{t+1})|x_{0:t}], \quad \Omega_{N,t} = \sum_{i=1}^{N} w_i^t. \quad (25) \]

Appendix B further connects our model and algorithm to the popular GHK sampler, hidden Markov models (HMMs), and PF and SMC techniques.

The SIS algorithm has a fundamental weakness called “weight degeneracy”: as the algorithm propagates through an increasing number of iterations, a large number of the normalized weights become negligible. As a result, only a few particles “contribute” in the likelihood approximation. Following the developments in the SMC (see [14], [31] and [8]) and HMM literatures (Sections 10.4.1 and 10.4.2 in [13]), we modify the SIS algorithm by adding a resampling step (all future simulations and computations use resampling).

**Sequential Importance Sampling with Resampling (SISR):** Proceed as in the SIS algorithm, but modify Step 2 and add a resampling Step 3 as follows:
2: Modify Step 2 of the SIS by setting

\[
\tilde{Z}^i_t \sim \mathcal{N}_{A_{x_t}}(\tilde{Z}^i_t, r^2_t), \quad \tilde{w}_t^i = w_{t-1}^i w_t(\tilde{Z}^i_t), \quad \tilde{\Omega}_{N,t} = \sum_{i=1}^{N} \tilde{w}_t^i.
\]  

(26)

3: For each particle \(i \in \{1, \ldots, N\}\), draw, conditionally and independently given \(\{(Z^i_s, w^i_s), s \leq t-1\}\) and \(\{\tilde{Z}^i_t\}\), a multinomial trial \(I^i_t \sim \{0, 1\}\) for each \(t\) and \(i\) with the success probabilities \(\{\tilde{w}_t^i / \tilde{\Omega}_{N,t}\}\) and set \(Z^i_t = I^i_t \tilde{Z}^i_t\) and \(w^i_t = 1\).

While the resampling step removes particles with low weights, mitigating degeneracy issues, it introduces additional estimator variance. We follow standard practice and resample only when the variance of the weights exceeds a certain threshold, quantified by the so-called effective sample size defined as \(\text{ESS}(w^i_t) = (\sum_{i=1}^{N} (w^i_t / \tilde{\Omega}_{N,t})^2)^{-1}\), and the resampling step is executed when \(\text{ESS}(w^i_t) < N/2\) as in [15]. See also Section 2.5.3 in [31] for a justification of the ESS based on the Delta method.

3 Inference

The model in (1) contains the parameters \(\theta\) in the marginal count distribution \(F_X\) and \(\eta\) in the dependence structure of \(\{Z_t\}\). This section addresses inference questions, including parameter estimation and goodness-of-fit assessment. Three methods are presented for parameter estimation: Gaussian pseudo-likelihood, implied Yule-Walker moment methods, and full likelihood. Gaussian pseudo-likelihood estimators, a time series staple, pretend that the series is Gaussian and maximize its Gaussian-based likelihood. These estimators only involve the mean and covariance structure of the series, are easy to compute, and will provide a comparative basis for likelihood estimators. They can also be used as initial
guesses in gradient step-and-search likelihood optimizations. Implied Yule-Walker techniques are moment based estimators applicable to the commonly encountered case where \( \{Z_t\} \) is a causal autoregression. Likelihood estimators, the statistical gold standard and the generally preferred estimation technique, are based on the PF and SMC methods of the last section. Finally, we will not delve into a detailed statistical inference for the aforementioned methods: while consistency and asymptotic normality are expected in some of the examined cases (e.g. likelihood estimation with an autoregressive \( \{Z_t\} \)), a rigorous theoretical treatment is beyond the scope of this paper.

3.1 Gaussian pseudo-likelihood estimation

As in Section 2.3, we work with observations \( x_t \) for the times \( t \in \{0, \ldots, T\} \) and set \( X = (x_0, \ldots, x_T)' \). Denote the likelihood of the model in (1) by

\[
\mathcal{L}_T(\theta, \eta) = \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_T = x_T) = p(x_{0:T}).
\]

While this likelihood is a multivariate normal probability, it is difficult to calculate or approximate when \( T \) is large. For most count model classes, true likelihood estimation is difficult to conduct as joint distributions are generally intractable [11]. While Section 3.3 below devises a well performing PF/SMC likelihood approximation (see also [39]), we first consider a simple Gaussian pseudo-likelihood (GL) approach. In a pseudo GL approach, parameters are estimated via

\[
(\hat{\theta}, \hat{\eta}) = \arg\max_{\theta, \eta} e^{-\frac{1}{2}(X - \mu_{\theta})' \Gamma_T(\theta, \eta)^{-1} (X - \mu_{\theta})}
\]

\[
\cdot \frac{1}{(2\pi)^{(T+1)/2} |\Gamma_T(\theta, \eta)|^{1/2}},
\]

where \( \mu_{\theta} = (\mu_{\theta}, \ldots, \mu_{\theta})' \) is a \((T + 1)\)-dimensional constant mean vector. These estimators maximize the series’ likelihood assuming the data are Gaussian, each component having
mean $\mu_\theta$, and all components having covariance matrix $\Gamma_T(\theta, \eta) = (\gamma_X(i - j))_{i,j=0}^T$. Time series analysts have been maximizing Gaussian pseudo likelihoods for decades, regardless of the series’ marginal distribution, with often satisfactory performance. The next section and Appendix C present a case where this approach works reasonably well, and one where it does not. For large $T$, the pseudo GL approach is equivalent to least squares estimation, where the sum of squares $\sum_{t=0}^T (X_t - \mathbb{E}[X_t|X_0, \ldots, X_{t-1}])^2$ is minimized (see Chapter 8 in [6]). The covariance structure of $\{X_t\}$ was efficiently computed in Section 2; the mean $\mu_\theta$ is usually explicitly obtained from the marginal distribution $F_X$ posited. Numerical optimization of (28) yields a Hessian matrix that can be inverted to obtain standard errors for the model parameters. These standard errors can be asymptotically corrected for distributional misspecification via the sandwich methods of [19].

### 3.2 Implied Yule-Walker estimation for latent AR models

Suppose that $\{Z_t\}$ follows the causal AR($p$) model $Z_t = \phi_1Z_{t-1} + \ldots + \phi_pZ_{t-p} + \varepsilon_t$, where $\{\varepsilon_t\}$ consists of IID $\mathcal{N}(0,\sigma_\varepsilon^2)$ variables. Here, $\sigma_\varepsilon^2$ depends on the autoregressive coefficients $\phi_1, \ldots, \phi_p$ in a way that induces $\mathbb{E}[Z_t^2] = 1$. The Yule-Walker equations are

$$\phi = \Gamma_p^{-1}\gamma_p,$$

(29)

where $\Gamma_p = (\gamma_Z(i - j))_{i,j=1}^p$, $\gamma_p = (\gamma_Z(1), \ldots, \gamma_Z(p))'$, and $\phi = (\phi_1, \ldots, \phi_p)'$. From (6), note that

$$\gamma_Z(h) = L^{-1}(\rho_X(h)),$$

(30)

the inverse being justified via the strictly increasing nature of $L(u)$ in $u$.

Equations (29) and (30) suggest the following estimation procedure. First, estimate the CDF parameter $\theta$ directly from the counts; standard methods (e.g. method of moments) are
typically available for this task. The estimated parameter $\hat{\theta}$ defines an estimated link $\hat{L}(u)$ through its estimated power series coefficients. From a numerical power series reversion procedure, one can now efficiently construct the inverse estimator $\hat{L}^{-1}(\rho)$.

Next, in view of (30) and (29), set

$$\hat{\gamma}_Z(h) = \hat{L}^{-1}(\hat{\rho}_X(h)), \quad \hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p,$$

where $\hat{\rho}_X(h)$ is the lag-$h$ sample autocorrelation of $\{X_t\}$, and $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ are defined analogously to the above using $\hat{\gamma}_Z(h)$ in place of $\gamma_Z(h)$.

### 3.3 Particle filtering and sequential Monte Carlo likelihoods

Using (23) and its notation, the true likelihood in (27) is

$$L_T(\theta, \eta) = p(x_0) \prod_{s=1}^{T} p(x_s|x_{0:s-1}) = p(x_0) \prod_{s=1}^{T} \mathbb{E}[1_{\{x_s\}}(X_s)|x_{0:s-1}] = p(x_0) \prod_{s=1}^{T} \mathbb{E}[w_s(\hat{Z}_s)|x_{0:s-1}],$$

where (23) was used with $D_{1(x_s),s}(z) = w_s(z)$ and $w_s(z)$ is defined and numerically computed akin to (19). The particle approximation of the likelihood is then

$$\hat{L}_T(\theta, \eta) = p(x_0) \prod_{s=1}^{T} \mathbb{E}[w_s(\hat{Z}_s)|x_{0:s-1}];$$

this uses the notation in (13) and supposes that the particles are generated by one of the methods in Section 2.3. The approximate PF maximum likelihood estimates satisfy

$$(\hat{\theta}, \hat{\eta}) = \arg\max_{\theta, \eta} \hat{L}_T(\theta, \eta).$$

**Remark 3.1.** With the SIS algorithm, (33) reduces to

$$\hat{L}_T(\theta, \eta) = p(x_0) \frac{1}{N} \sum_{i=1}^{N} w_T^i,$$

19
which is consistent with (22). The work [35] also essentially implements (35). In contrast to [35], our approach includes a resampling step in the likelihood approximations, considers other estimation approaches (pseudo GL and implied Yule-Walker), and provides model diagnostic tools more specific to count series (the PIT histograms in Section 3.4 below).

To optimize the estimate $\hat{L}_T(\theta, \eta)$, we employ a large number of particles (growing linearly with $T$) and common random number (CRN) techniques, a standard practice that serves to smooth $\hat{L}_T(\theta, \eta)$ somewhat by expressing its random quantities through parameter-dependent transformations of uniform random variables that remain constant for likelihood evaluations across distinct parameters. While the CRN procedure works well in SIS, it fails to ward against discontinuous $\hat{L}_T(\theta, \eta)$ in our preferred SISR algorithm. An elegant solution to this issue for univariate state processes is proposed in [34]: first reorder the (real-valued) particles and then replace the discontinuous resampling CDF with a piecewise linear approximation. More recent and well performing (but less straightforward) approaches such as the sequential quasi Monte Carlo and the SMC$^2$ algorithm are reviewed in detail in Chapters 13, 14, and 18 of [8] (see also the Chapter 19 references on controlled sequential Monte Carlo methods). We do not pursue these issues further here.

In our numerical implementations, gradient-free algorithms from the R package optimx [36] are used, which follows standard practices in optimizing noisy objective functions. These routines allow for boundary constraints and performed well in modest computing times for our sample sizes. On the other hand, we found less success with the more popular gradient-based quasi-Newton algorithm L-BFGS-B (gradients were computed via finite differences) as convergence instabilities and high-variance estimates were encountered. However, promising recent developments for optimizing noisy objectives in [4] and [38] were not explored. A comprehensive investigation of these approaches and of the rich gradient-
based SMC inference literature for our framework as in [27] is deferred to future work.

3.4 Model diagnostics

The goodness-of-fit of count models is commonly assessed through probability integral transform (PIT) histograms and related tools [10, 29]. These are based on the predictive distributions of \( \{X_t\} \), defined at time \( t \) by

\[
P_t(y) = \mathbb{P}(X_t \leq y | X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t \leq y | x_0, \ldots, x_{t-1}), \quad y \in \{0, 1, \ldots\}. \tag{36}
\]

This quantity can be estimated through the PF/SMC methods in Section 2.3 as

\[
\hat{P}_t(y) = \sum_{\ell=0}^{y} \hat{E}[1_{\{\ell\}}(X_t)|x_0:t-1] = \sum_{\ell=0}^{y} \hat{E}[D_{1(\ell),t}(\hat{Z}_t)|x_0:t-1], \tag{37}
\]

which uses (24) and (25) and supposes that the particles are generated by the SIS, SISR, or other algorithms. Similar to \( D_{1(x_s),s}(z) = w_s(z) \), note that \( D_{1(x),t}(z) = \tilde{w}_{x,t}(z) \), where

\[
\tilde{w}_{x,t}(z) = \Phi \left( \frac{\Phi^{-1}(C_x) - z}{r_t} \right) - \Phi \left( \frac{\Phi^{-1}(C_{x-1}) - z}{r_t} \right) \tag{38}
\]

and \( \tilde{w}_{x,t}(z) = w_t(z) \).

The (non-randomized) sample mean PIT is defined as

\[
\overline{F}(u) = \frac{1}{T+1} \sum_{t=0}^{T} F_t(u|x_t), \quad u \in [0, 1], \tag{39}
\]

where

\[
F_t(u|y) = \begin{cases} 
0, & \text{if } u \leq P_t(y-1), \\
\frac{u - P_t(y-1)}{P_t(y) - P_t(y-1)}, & \text{if } P_t(y-1) < u < P_t(y), \\
1, & \text{if } u \geq P_t(y),
\end{cases} \tag{40}
\]
which is estimated by replacing $P_t$ by $\hat{P}_t$ in practice. The PIT histogram with $H$ bins is defined as a histogram with the height $\overline{F}(h/H) - \overline{F}((h - 1)/H)$ for bin $h \in \{1, \ldots, H\}$.

Another possibility considers model residuals based on

$$
\hat{Z}_t = \mathbb{E}[Z_t|X_t = x_t] = \frac{\exp(-\Phi^{-1}(C_{x_t-1})^2/2) - \exp(-\Phi^{-1}(C_{x_t})^2/2)}{\sqrt{2\pi(C_{x_t} - C_{x_t-1})}},
$$

which is the estimated mean of the latent Gaussian process at time $t$ given $X_t$ only (not the entire past), where (41) follows by direct calculations for the model (1) (using the estimated parameters $\theta$ of the marginal distribution in the $C_k$s). For a fitted underlying time series model with parameter $\eta$, the residuals are then defined as the standard time series residuals $\hat{\epsilon}_t$ of this model fitted to the series $\hat{Z}_t$, after centering by the sample mean.

### 3.5 Nonstationarity and covariates

As discussed in Section 2.2, covariates can be accommodated by allowing a time-varying parameter $\theta$ in the marginal distribution. With covariates, $\theta$ at time $t$ is denoted by $\theta(t)$. The GL and PF/SMC procedures are modified for $\theta(t)$ as follows.

For the GL procedure, the covariance $\text{Cov}(X_{t_1}, X_{t_2}) = \text{Cov}(G_{\theta(t_1)}(Z_{t_1}), G_{\theta(t_2)}(Z_{t_2}))$ is needed, where $G$ is subscripted to signify dependence on $\theta$. But as in (5),

$$
\text{Cov}(X_{t_1}, X_{t_2}) = \text{Cov}(G_{\theta(t_1)}(Z_{t_1}), G_{\theta(t_2)}(Z_{t_2})) = \sum_{k=1}^{\infty} k! g_{\theta(t_1), k} g_{\theta(t_2), k} \gamma Z(t_1 - t_2)^k,
$$

where again, the subscript $\theta(t)$ is added to the $g_k$s to indicate dependence on $t$. Numerically, evaluating (42) is akin to the task in (5); in particular, both calculations are based on the Hermite coefficients $\{g_k\}$.

For the PF/SMC approach, the modification is somewhat simpler: one just needs to replace $\theta$ by $\theta(t)$ at time $t$ when generating the underlying particles. For example, for the
SIS algorithm, \( \theta(t) \) enters only through the \( C_x \)'s in (19), (20), and (21). This is because the covariates enter only through \( \theta \), the parameter controlling marginal distributions.

4 A simulation study

To evaluate our estimation methods, a simulation study considering several marginal distributions and dependence structures was conducted. Here, the classic Poisson count distribution \( P \) is examined (mixed Poisson and negative binomial simulations are presented in Appendix C), with \( \{Z_t\} \) taken from the ARMA\((p,q)\) class. All simulation cases are replicated 200 times for three distinct series lengths: \( T = 100, 200, \) and 400. For notation, estimates of a parameter \( \zeta \) from Gaussian pseudo-likelihood (GL), implied Yule-Walker (IYW), and PF/SMC methods are denoted by \( \hat{\zeta}_{GL} \), \( \hat{\zeta}_{IYW} \), and \( \hat{\zeta}_{PF} \), respectively.

We now consider the classical case where \( X_t \) has a Poisson marginal distribution for each \( t \) with mean \( \lambda > 0 \). To obtain \( X_t \), the AR(1) process \( Z_t = \phi Z_{t-1} + (1 - \phi^2)^{1/2} \epsilon_t \), was simulated and transformed via (1) with \( F = P; \ E[Z_t^2] \equiv 1 \) was induced by taking \( \text{Var}(\epsilon_t) \equiv 1 \). Twelve parameter schemes resulting from all combinations of \( \lambda \in \{2, 5, 10\} \) and \( \phi \in \{\pm 0.25, \pm 0.75\} \) were considered.

Figure 1 displays box plots of the parameter estimates when \( \lambda = 2 \). In estimating \( \lambda \), all methods perform reasonably well. When the lag-one correlation in \( \{Z_t\} \) (and hence also that in \( \{X_t\} \)) is negative (right panel), \( \hat{\lambda}_{GL}, \hat{\lambda}_{IYW}, \) and \( \hat{\lambda}_{PF} \) have smaller variability than the positively correlated case (left panel — note the different y-axis scales on the panels). This is expected: the mean of \( \{X_t\} \) is \( \lambda \), and the variability of the sample mean, one good estimator of the mean for a stationary series, is smaller for negatively correlated series than for positively correlated ones. Note that \( \hat{\phi}_{GL} \) is biased toward zero for both negatively and
positively correlated series, whereas \( \hat{\phi}_{IW} \) and \( \hat{\phi}_{PF} \) only show bias when \( \phi \) is positive for the sample sizes \( T = 100 \) and \( T = 200 \). Overall, the PF/SMC estimates were the least biased. All estimates of \( \phi \) have roughly similar variances. Simulations with \( \lambda = 5 \) and \( \lambda = 10 \) produced analogous results with smaller values of \( \lambda \) yielding less variable estimates. This is again expected as the variance of the Poisson distribution is also \( \lambda \). Graphics of these box plots are omitted for brevity’s sake.

Figure 1: Gaussian likelihood, implied Yule-Walker, and PF/SMC parameter estimates for 200 synthetic Poisson–AR(1) series with lengths \( T = 100, 200, \) and \( 400 \). The true parameter values (indicated by horizontal dashed lines) are \( \lambda = 2 \) and \( \phi = 0.75 \) (left panel), and \( \lambda = 2 \) and \( \phi = -0.75 \) (right panel).

5 An application

This section applies our methods to a weekly count series of product sales at Dominicks Finer Foods, a now defunct U.S. grocery chain that operated in Chicago, IL and adjacent areas
from 1918 - 2013. Soft drink sales of an unnamed brand from a single store will be analyzed
over a two-year span commencing on September 10, 1989. The series is plotted in Figure
2 (leftmost plot) and is part of a large and well-studied retail dataset, publicly available at
https://www.chicagobooth.edu/research/kilts/datasets/dominicks (Source: The James M.
Kilts Center for Marketing, University of Chicago).\textsuperscript{1} Our goal here is not an in-depth retail

\footnotesize
\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Left: Weekly sales of a soft drink product sold at a single store of the grocery store
Dominick’s Finer Foods from 09-10-1989 to 09-03-1991. The dots indicate the weekly sales were
at least one “Buy one and get one free” (BOGO) sales promotion event took place. Middle:
Boxplots of sales grouped by the BOGO covariate (0: weekly sales with no BOGO event, 1: weekly
sales with at least one BOGO day during the week). Right: Sample ACF of the series with 95%
pointwise bands for zero correlation.}
\end{figure}

analysis, but to illustrate our methods with a real world example of an overdispersed time
series of small counts that has negative autocorrelation and dependence on a covariate.

The covariate we use is a zero-one “buy one get one free” (BOGO for short) sales
promotion event $S_t$, $S_t = 1$ implying that the BOGO promotion was offered at least one
day during week $t$. The dots in the left plot of Figure 2 signify that the week had at

\footnotesize
\textsuperscript{1}In the dataset manual, the series in Figure 2 (leftmost plot) is the sales of the product with universal
product code (UPC) 4640055081 from store 81.
least one BOGO day. The middle plot shows the soft drinks sales distribution grouped by \( S_t \), visually suggesting that a BOGO event increases soft drink sales. The rightmost plot shows the sample ACF of the series and reveals negative dependence at lag one. The lag one sample autocorrelation of the residuals after a linear regression of the series on the BOGO covariate is also negative, but comparatively smaller in magnitude.

To model overdispersion, negative binomial and generalized Poisson marginal distributions will be considered. Although similar, these two distributions can yield different conclusions [25]. Following standard generalized linear modeling practice, both distributions are parametrized via the series’ mean (although our setup allows covariates to enter through other parameters as well). More specifically, for the negative binomial marginal, the standard pair \((r, p)\) used in Appendix C is now mapped to the parameter pair \((\mu, k)\), where \( \mu = pr/(1 - p) \) is the mean of the process and \( k = 1/r \) is the overdispersion parameter. Similarly, the generalized Poisson distribution of Appendix C is parametrized through the pair \((\mu, \alpha)\) as in [17], relation (2.4). In this parametrization, \( \mu \) is the mean of the series, whereas the sign of \( \alpha \) controls the type of dispersion, with positive values indicating overdispersion. To incorporate the BOGO covariate \( S_t \) into the model, the mean of the series is allowed to depend on time \( t \) through the typical GLM log-link \( \mu_t = \exp(\beta_0 + \beta_1 S_t) \), while the parameters \( k \) and \( \alpha \) are kept fixed in time \( t \).

An exploratory examination of the sample ACFs and PACFs of the series along with diagnostic plots of residuals obtained by fitting all ARMA\((p, q)\) models with \( p, q \leq 5 \) suggest an AR(3) model as a suitable choice for \( \{Z_t\} \). Table 1 in Appendix D shows the AICc and BIC for both marginal distributions obtained via PF/SMC and GL methods (we omit IYW results for simplicity). The AR(3) model was selected by AICc and BIC in both fits. Interestingly, both the sample ACF and PACF of the series show one large non-zero value.
at lag one, but relatively smaller values at other lags (except perhaps the lag two value, which barely exceeds the 95% $1.96/\sqrt{T}$ dashed confidence threshold for zero correlation).

We also considered a white noise latent series (labeled as “WN” Table 1 in Appendix D), which renders our model a standard GLM. The PF/SMC WN estimates from both distributions (omitted here for brevity) closely agree with parameter estimates obtained from exact generalized linear model fits (using, for example, functions from the R package “MASS”). As expected, the WN model yielded the highest AICc and BIC values among all considered dependence structures, thus confirming the need for a model with temporal dependence.

Table 1 shows parameter estimates and standard errors from fitting a negative binomial-AR(3). (Table 2 in Appendix D is for a generalized Poisson-AR(3) model.) All marginal distributions and estimation methods yielded $\hat{\phi}_1 < 0$. Although a formal asymptotic theory is beyond the scope of our presentation here, asymptotic normality is expected. Assuming this, the PF/SMC standard errors (the ones believed most trustworthy) suggest that all parameters are significantly non-zero at level 95%. The findings suggest the negative binomial distribution is preferred over the generalized Poisson, that the correlation in the series at lag one is negative, and that a BOGO event indeed increases sales.

| Parameters                  | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\beta_0$ | $\beta_1$ | $k$  |
|-----------------------------|----------|----------|----------|-----------|-----------|------|
| GL Estimates                | -0.447   | 0.145    | 0.208    | 2.433     | 0.569     | 0.884|
| GL Standard Errors          | 0.175    | 0.171    | 0.130    | 0.095     | 0.115     | 0.207|
| PF/SMC Estimates            | -0.341   | 0.223    | 0.291    | 2.264     | 1.01      | 1.21 |
| PF/SMC Standard Errors      | 0.100    | 0.107    | 0.102    | 0.142     | 0.207     | 0.205|

Table 1: Estimates and standard errors of the negative binomial-AR(3) model.

Turning to residual diagnostics, the plots in Figure 3 for the negative binomial-AR(3)
fit suggest that the model has captured both the marginal distribution and the dependence structure. The residuals here were computed using (41).

Figure 3: The leftmost plot displays the estimated residuals against time. The middle graph is a QQ plot for normality of the estimated residuals. The shaded region in the QQ plot shows 100 realizations from a normal distribution with size, mean and standard deviation matching the residual sample counterparts. The right plot displays the sample autocorrelations of the estimated residuals.

We next assess the predictive ability of the two fits via the non-randomized histograms shown in Figure 4 and discussed in detail in Section 3.4. We selected ten bins at the points \( h/10, h = 1, \ldots, 10 \) as is typical in the literature. The negative binomial PIT plot suggests a satisfactory predictive ability with most bar heights being close to 0.1 (1 over the number of bins). In comparison, the generalized Poisson fit deviates more from the uniform distribution, with somewhat more pronounced peaks and valleys. We remind the reader here that PIT plots are known to be sensitive for smaller series lengths. Quantifying this uncertainty (for each bin) through a statistical test is beyond the scope of this paper. Nevertheless, we gauged the variability of the uniform distribution’s bin heights through a small experiment. Specifically, 500 synthetic realizations of sample size \( T = 104 \) were generated and the percentiles of all bin heights were collected. The 5th and 95th percentiles
ranged in the intervals (0.048, 0.058) and (0.145, 0.154) respectively, suggesting that the peaks and valleys of the negative binomial PIT plot (which are within these percentiles) are mild; that is, uniformity is plausible and the marginal distribution fits seems adequate.

Figure 4: PIT residual histograms for the estimated models in Tables 1 and 2.

6 Conclusions and comments

This paper developed the theory and methods for a stationary count time series model made from a latent Gaussian process. By using Hermite expansions, a very general model class was devised. In particular, the autocorrelations in the series can be positive or negative, and in a pairwise sense, span the range of all achievable correlations. The series can have any marginal distribution desired, thereby improving classical DARMA and INARMA count time series methods. On inferential levels, autocovariances of the model were extracted from Hermite expansions, allowing for Gaussian pseudo-likelihood and implied Yule-Walker inference procedures. A PF/SMC likelihood approach was also developed and produced estimators that were demonstrated to outperform the Gaussian
pseudo-likelihood and implied Yule-Walker estimators in most cases. These results complement the importance sampling methods for copula likelihoods in [39]. The methods were used in a simulation study and were applied in a regression analysis of a count series of weekly grocery sales that exhibited overdispersion, a negative lag one correlation, and dependence on a “buy one get one free” covariate. Model fits and predictive abilities of the methods were illustrated with generalized Poisson and negative binomial marginal distributions.

While the paper provides a reasonably complete treatment for count time series models, additional research is needed. Some statistical issues, like asymptotic normality of parameter estimators, were not addressed here. PF/SMC algorithms that optimize model likelihoods, which can be unwieldy, also merit further exploration. The paper only considers univariate methods. Multivariate count time series models akin to those in [40] could be developed by replacing the univariate \( \{Z_t\} \) with a multivariate Gaussian process \( \{Z_t\} \), whose components have a standard normal marginal distribution, but are cross-correlated for each fixed \( t \). The details for such a construction would proceed akin to the methods developed here. Also, while the count case is considered here, the same methods will produce stationary time series having any general prescribed continuous distribution. Finally, the same methods should prove useful in constructing spatial and spatio-temporal processes having any prescribed marginal distribution. While [12, 21] recently addressed this issue in the spatial setting, additional work is needed, including exploring spatial Markov properties and likelihood evaluation techniques. To the best of our knowledge, no comprehensive analogous work has been conducted for space-time count modeling to date.
References

[1] Asmussen, S. (2014). Modeling and performance of bonus-malus systems: stationarity versus age-correction. *Risks* 2, 49–73.

[2] Belyaev, M., E. Burnaev, and Y. Kapushev (2015). Gaussian process regression for structured data sets. In A. Gammerman, V. Vovk, and H. Papadopoulos (Eds.), *Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015*. Switzerland: Springer International Publishing.

[3] Benjamin, M. A., R. A. Rigby, and D. M. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association* 98, 214–223.

[4] Berahas, A. S., R. H. Byrd, and J. Nocedal (2019). Derivative-free optimization of noisy functions via quasi-Newton methods. *SIAM Journal on Optimization* 29, 965–993.

[5] Blight, P. A. (1989). Time series formed from the superposition of discrete renewal processes. *Journal of Applied Probability* 26, 189–195.

[6] Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). New York City: Springer-Verlag.

[7] Chen, H. (2001). Initialization of NORTA: Generation of random vectors with specified marginals and correlations. *Inform Journal on Computing* 13, 312–331.

[8] Chopin, N. and O. Papaspiliopoulos (2020). *An Introduction to Sequential Monte Carlo Methods*. New York City: Springer.
[9] Cui, Y. and R. B. Lund (2009). A new look at time series of counts. *Biometrika* 96, 781–792.

[10] Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics* 65, 1254–1261.

[11] Davis, R. A., S. H. Holan, R. B. Lund, and N. Ravishanker (Eds.) (2016). *Handbook of Discrete-Valued Time Series*. Boca Raton, Florida, USA: CRC Press.

[12] De Oliveira, V. (2016). Hierarchical Poisson models for spatial count data. *Journal of Multivariate Analysis* 122, 393–408.

[13] Douc, R., E. Moulines, and D. S. Stoffer (2014). *Nonlinear Time Series: Theory, Methods, and Applications with R Examples*. Boca Raton, Florida, USA: CRC Press.

[14] Doucet, A., N. De Freitas, and N. Gordon (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*, pp. 3–14. New York City: Springer.

[15] Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering* 12, 656–704.

[16] Dumsmuir, W. T. M. (2016). Generalized linear autoregressive moving-average models. In *Handbook of Discrete-valued Time Series*, pp. 51–76. Boca Raton, Florida, USA: CRC Press.

[17] Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics-Theory and Methods* 22, 1335–1354.
[18] Fokianos, K. (2012). Count time series models. In *Handbook of Statistics*, Volume 30, pp. 315–347. Amsterdam: Elsevier.

[19] Freedman, D. (2006). On the so-called “Huber sandwich estimator” and “robust standard errors”. *The American Statistician* 60, 299–302.

[20] Grigoriu, M. (2007). Multivariate distributions with specified marginals: applications to wind engineering. *Journal of Engineering Mechanics* 133, 174–184.

[21] Han, Z. and V. De Oliveira (2016). On the correlation structure of Gaussian copula models for geostatistical count data. *Australian & New Zealand Journal of Statistics* 58, 47–69.

[22] Han, Z. and V. De Oliveira (2020). Maximum likelihood estimation of Gaussian copula models for geostatistical count data. *Communications in Statistics - Simulation and Computation* 49, 1957–1981.

[23] Jacobs, P. A. and P. A. W. Lewis (1978a). Discrete time series generated by mixtures I: Correlational and runs properties. *Journal of the Royal Statistical Society* 40, 94–105.

[24] Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability* 33, 664–677.

[25] Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47, 219–229.

[26] Kachour, M. and J. F. Yao (2009). First order rounded integer valued autoregressive (RINAR(1)) processes. *Journal of Time Series Analysis* 30, 417–448.
[27] Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* 30, 328–351.

[28] Kedem, B. (1980). Estimation of the parameters in stationary autoregressive processes after hard limiting. *Journal of the American Statistical Association* 75, 146–153.

[29] Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting* 32, 788–803.

[30] Lennon, H. (2016). *Gaussian copula modelling for integer-valued time series*. Ph. D. thesis, The University of Manchester.

[31] Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. New York City: Springer Science & Business Media.

[32] Liu, J. S. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association* 93, 1032–1044.

[33] Livsey, J., R. B. Lund, S. Kechagias, and V. Pipiras (2018). Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts. *Annals of Applied Statistics* 12, 408–431.

[34] Malik, S. and M. K. Pitt (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics* 165, 190–209.

[35] Masarotto, G. and C. Varin (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* 6, 1517–1549.

[36] Nash, J. C. and R. Varadhan (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software* 43, 1–14.
[37] Pipiras, V. and M. S. Taqqu (2017). *Long-Range Dependence and Self-Similarity*, Volume 45. Boca Raton, Florida, USA: Cambridge University Press.

[38] Shi, H.-J. M., M. Q. Xuan, F. Oztoprak, and J. Nocedal (2021). On the numerical performance of derivative-free optimization methods based on finite-difference approximations. *arXiv preprint arXiv:2102.09762*.

[39] Smith, M. S. and M. A. Khaled (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association 107*, 290–303.

[40] Song, P., M. Li, and P. Zhang (2013). Vector generalized linear models: a Gaussian copula approach. In P. Jaworski, F. Durante, and W. Härdle (Eds.), *Copulae in Mathematical and Quantitative Finance*. Heidelberg, Germany: Springer.

[41] Tong, Y. L. (1990). *The Multivariate Normal Distribution*. New York City: Springer-Verlag.

[42] Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of Statistics 4*, 1280–1289.

[43] Zheng, T., H. Xiao, and R. Chen (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics 189*, 492–506.