The open-faced sandwich adjustment for MCMC using estimating functions

Benjamin A. Shaby
Department of Statistics
University of California, Berkeley
429 Evans Hall, Berkeley, CA 94720, U.S.A.

Abstract

The situation frequently arises where working with the likelihood function is problematic. This can happen for several reasons—perhaps the likelihood is prohibitively computationally expensive, perhaps it lacks some robustness property, or perhaps it is simply not known for the model under consideration. In these cases, it is often possible to specify alternative functions of the parameters and the data that can be maximized to obtain asymptotically normal estimates. However, these scenarios present obvious problems if one is interested in applying Bayesian techniques. This article describes open-faced sandwich adjustment, a way to incorporate a wide class of non-likelihood objective functions within Bayesian-like models to obtain asymptotically valid parameter estimates and inference via MCMC. Two simulation examples show that the method provides accurate frequentist uncertainty estimates. The open-faced sandwich adjustment is applied to a Poisson spatio-temporal model to analyze an ornithology dataset from the citizen science initiative eBird. An online supplement contains an appendix with additional figures, tables, and discussion, as well as R code.

Keywords: covariance tapering, max-stable process, calibration, estimating functions
1 Introduction

For many models arising in various fields of statistical analysis, working with the likelihood function can be undesirable. This may be the case for several reasons—perhaps the likelihood is prohibitively expensive to compute, perhaps it presumes knowledge of a component of the model that one is unwilling to specify, or perhaps its form is not even known for a chosen probability model. Such scenarios present problems if one wishes to perform Bayesian analysis. Applying the Bayesian computational and inferential machinery, thereby enjoying benefits such as natural shrinkage, variance propagation, and the ability to incorporate complex hierarchical dependences, usually requires working directly with the likelihood function.

To motivate development, I briefly describe the analysis of bird sightings contained in Section 5. The data consist of several thousand counts occurring irregularly in space and time (see Figure S-3), along with several spatially-varying covariates carefully chosen by a group of ornithologists. The analysis has two related goals. The first is to produce maps of bird abundance, along with associated uncertainties, to inform conservation measures. Second, the ornithologists are interested in the spatio-temporal structure of the abundance surface that is not explained by the covariates. This structure gives a hint about what kinds of additional covariates are useful to collect, as well as some suggestion about the spatio-temporal aspects of the bird populations themselves. A natural model for such abundance data is a hierarchical Poisson regression with a random effect specified as a spatio-temporal Gaussian process with unknown covariance parameters. For this model, maximum likelihood and related methods are not possible because of the intractability of integrating out the high-dimensional spatio-temporal random effect. Bayesian methods are therefore the tool of choice based on computational necessity, and in addition provide sharing of information across space and time, as well as automatic uncertainty estimation of predictive abundance maps. Furthermore, obtaining an MCMC sample of the posterior distribution is desirable because inferences on the posterior correlation surface of the random effect, a nonlinear functional of random covariance parameters, is of independent interest to the ornithologists. However, the sheer size of the dataset makes MCMC under this model intractable, so a faster objective function is used in place of the high-dimensional Gaussian likelihood. The goal of the method presented here is to enable such a substitution while retaining a valid interpretation of the resultant MCMC
sample.

More generally, suppose that one specifies a model, either one-stage or hierarchical, and wants the advantages of being Bayesian, but the likelihood in some level of the hierarchy is problematic. Suppose, however, that one can write down some objective function $\ell_M(\theta; y)$ of the parameters and the data (possibly conditional on other parameters) that behaves similarly to the log likelihood. What is meant by “similarly” is defined in Section 2. Important examples of methods that employ such objective functions include generalized method of moments (Hall et al. 2005) and robust M-estimation (Huber and Ronchetti 2009), as well as the two examples we will consider here, covariance tapering (Kaufman et al. 2008) and composite likelihoods (Lindsay 1988).

The question that the present work attempts to answer here is this: Can one insert $\ell_M(\theta; y)$, in place of the likelihood, into an MCMC algorithm like Metropolis-Hastings and “trick” it into doing something useful? I claim that one can—that for many useful examples, simply swapping $\ell_M(\theta; y)$ into a sampler results in a quasi-posterior sample that can be rotated and scaled to yield desirable properties.

The open-faced sandwich (OFS) adjustment relies on asymptotic theory that was formally developed in Chernozhukov and Hong (2003), but which is quite intuitive. These authors were interested in using Metropolis-Hastings as an optimization algorithm for badly-behaved objective functions, not in using non-likelihood objective functions for performing Bayesian-like analysis, as is the case here. Although their goals were entirely different, the theory contained therein is extremely useful for the present purposes.

Previous attempts to incorporate non-likelihood objective functions into the Bayesian setting, to my knowledge, have been few. McVean et al. (2004) use composite likelihoods within reversible jump MCMC, without any adjustment, to estimate population genetic parameters. Realizing that their sampler would result in invalid inferences, McVean et al. (2004) turn to a parametric bootstrap to estimate sampling variability. Smith and Stephenson (2009) were interested in max-stable processes for spatial extreme value analysis. They also use composite likelihoods within MCMC without adjustment. The special case of using the generalized method of moments objective function (Hansen 1982; Hall 2005) for generalized linear models within an MCMC sampler was explored by Yin (2009). Tangentially related are Tian et al. (2007), who use MCMC to estimate the...
sampling distribution of $\hat{\theta} = \text{argmax} \ell_M(\theta; y)$, Müller (2009), who analyzes misspecified models using MCMC, and McCullagh and Tibshirani (1990), who adjust profile likelihoods to improve inference for models with nuisance parameters. Also related are the class of so-called approximate Bayesian computation algorithms (Sisson and Fan 2011, for a review), which simulate data conditional on proposed parameters, and accept or reject the proposal based on a comparison between a summary statistic of the conditional draw and a summary statistic of the observations.

Ribatet et al. (2012) attempt to solve the same problem that is addressed here. Whereas here quasi-posterior samples generated from MCMC are adjusted post hoc, these authors propose an adjustment to the Metropolis likelihood ratio within the sampler itself. Their goal, like the present one, is to achieve desirable frequentist coverage properties of credible intervals computed based on MCMC. Although their approach is quite general, Ribatet et al. (2012) restrict their attention to using composite likelihoods for max-stable processes. A related method was proposed by Pauli et al. (2011). The approach taken in the present article is closely related to that of Ribatet et al. (2012), but the OFS adjustment differs from their adjustment in its structure as well as its motivating asymptotic arguments.

Both the motivating insights for the OFS adjustment and the criterion by which it will be evaluated is essentially the idea of calibration (Rubin 1984; Draper 2006). In the present interpretation, a well-calibrated method has the property that when used to construct credible intervals from many different datasets, those intervals ought to cover the true parameter at close to their nominal rates. Essentially, this says that well-calibrated credible intervals behave like confidence intervals. If one constructs intervals with accurate coverage directly as the $\alpha/2$ and $(1 - \alpha/2)$ empirical quantiles of an MCMC sample for different values of $\alpha$, I claim that in some way the uncertainty about a parameter is well-described by the sample. Evaluating an approximate Bayesian method by this criterion has intuitive practical appeal, and it has been endorsed in particular by objective Bayesians (Bayarri and Berger 2004; Berger et al. 2001, e.g.).

This principle, along with some basic asymptotic observations, leads to the OFS adjustment. The asymptotic theory provides the limiting normal distribution of quasi-Bayes point estimators. One can take this distribution, in an informal sense, to be a summary of the uncertainty about $\theta$, up to an asymptotic approximation. The asymptotic theory also gives the limiting normal distribution
of the quasi-posterior. Since these two limiting distributions are not, in general, the same, and since one would like the quasi-posterior to summarize uncertainty about \( \theta \) in the sense of being well-calibrated, the strategy here is to adjust samples from the quasi-posterior so that their limiting distribution matches that of the quasi-Bayes point estimator.

There is a temptation to ask how well the adjusted quasi-posterior distribution approximates the true posterior distribution, in cases when the true likelihood is available. However, this is the incorrect comparison to make. The true posterior distribution contains the information about \( \theta \) obtained through the likelihood. When some other function \( \ell_M \) is used in place of the likelihood, there is no reason to expect the information content to remain the same. One would like the adjusted quasi-posterior distribution to represent this loss of information, not hide it. In the simulation examples in Section 4, the frequentist accuracy of credible intervals based on adjusted quasi-posterior samples shows that the OFS adjustment accomplishes this task.

Throughout, it will be assumed that expectations will be computed with respect to the true parameter \( \theta_0 \). Define the square root of a symmetric positive definite matrix \( A \) to be \( A^{1/2} = OD^{1/2}O' \), where \( A = ODO' \) with \( O \) orthogonal and \( D \) diagonal. The square root of a matrix is not unique; here \( A^{1/2} \) is computed using the singular value decomposition, which is numerically stable and preserves key geometric attributes.

Section 2 begins by defining the quasi-Bayesian framework and reviewing the relevant asymptotic theory. Section 3 develops the OFS adjustment method, and a demonstration of how to apply it in two different statistical contexts is contained in Section 4. In Section 5, the OFS adjustment is applied to analyze a dataset of Northern Cardinal sightings taken from the citizen science project eBird. Section 6 concludes.

## 2 The quasi-Bayesian framework

Begin by assuming that the parameter of interest \( \theta \) lies in the interior of a compact convex space \( \Theta \). Suppose we are given \( y \), which consists of \( n \) observations, from which we wish to estimate \( \theta \). Suppose further that we have at our disposal some objective function \( \ell_M(\theta; y) \) from which it is possible to compute \( \hat{\theta}_M = \arg\max_{\theta} \ell_M(\theta; y) \).
Following Chernozhukov and Hong (2003), define the \textit{quasi-posterior} distribution based on \( n \) observations \( y_n = y_1, \ldots, y_n \) as

\[
\pi_{M,n}(\theta|y_n) = \frac{L_{M,n}(\theta; y_n)\pi(\theta)}{\int_\Theta L_{M,n}(\theta; y_n)\pi(\theta) \, d\theta},
\]

where \( L_{M,n}(\theta; y_n) = \exp\{\ell_{M,n}(\theta; y_n)\} \), and \( \pi(\theta) \) is a prior density on \( \Theta \). It will be assumed, for convenience, that \( \pi(\theta) \) is proper with support on \( \Theta \). The function \( L_{M,n} \) is not necessarily a density, and thus \( \pi_{M,n}(\theta|y_n) \) is not a true posterior density in any probabilistic sense. It will be assumed, however, that \( L_{M,n} \) is integrable, so as long as the prior \( \pi(\theta) \) is proper, it easily follows that \( \pi_{M,n}(\theta|y_n) \) will be a proper density.

Equipped with notion of a quasi-posterior density, define quasi-posterior risk as

\[
R_n(\theta) = \int_\Theta \rho_n(\theta - \theta^*)\pi_{M,n}(\theta^*|y_n) \, d\theta^*,
\]

where \( \rho_n(u) \) is some symmetric convex scalar loss function. Then for a given loss function, the quasi-Bayes estimator is naturally defined as \( \hat{\theta}_{QB} = \arg\min_{\theta \in \Theta} R_n(\theta) \), the value of \( \theta \) that minimizes quasi-posterior risk.

The requirements on \( \ell_{M,n}(\theta; y_n) \) are fairly minimal and are met by most objective functions in wide use in statistics. Technical assumptions are contained in Chernozhukov and Hong (2003), but they are in general satisfied when \( \hat{\theta}_M \) is weakly consistent for \( \theta_0 \) and asymptotically normal.

Asymptotic normality of \( \hat{\theta}_M \) is of the form

\[
J_n^{1/2}(\hat{\theta}_M - \theta_0) \overset{D}{\to} N(0, I),
\]

where

\[
J_n = Q_n P_n^{-1} Q_n, \\
P_n = E_0[\nabla \ell_{M,n} \nabla \ell_{M,n}'], \\
Q_n = -E_0[\mathcal{H}_0 \ell_{M,n}].
\]

The notation \( \nabla f \) refers to the gradient of the function \( f \) evaluated at the true parameter \( \theta_0 \), and \( \mathcal{H} f \) refers to the Hessian of \( f \) evaluated at \( \theta_0 \). These matrices have been defined in terms of partial derivatives, but in general, \( \ell_{M,n} \) does not have to be differentiable or even continuous for the theory to apply. In this case, small adjustments of the definitions of \( P_n \) and \( Q_n \) are necessary (Chernozhukov and Hong 2003).
The sandwich matrix $J_n^{-1}$ is familiar from generalized estimating equations, quasi-likelihood, and other areas, and is referred to by various names, including the Godambe information criterion and the robust information criterion (e.g. Durbin 1960; Bhapkar 1972; Morton 1981; Ferreira 1982; Godambe and Heyde 1987; Heyde 1997). In the special case when $\ell_{M,n}(\theta; y)$ is the true likelihood, $Q_n \equiv J_n$, the Fisher information. It will hereafter be assumed that this is not the case.

Chernozhukov and Hong (2003) elucidates the asymptotic behavior of $\pi_{M,n}(\theta|y_n)$, which motivates the open-face sandwich adjustment. These results are direct analogues of well-known asymptotic properties of true posterior distributions. Here the statements of relevant results are kept deliberately imprecise because the assumptions in Chernozhukov and Hong (2003) are very general, technical, and lengthy, and they are not important for the present discussion. Their Theorem 2, which is re-stated below, states that the asymptotic distribution of the quasi-Bayes estimator $\hat{\theta}_{QB,n}$ is the same as that of the extremum estimator $\hat{\theta}_{M,n}$.

**Observation 1.** Assuming sufficient regularity of $\ell_{M,n}(\theta; y_n)$ and asymptotic normality of the form (2),

$$J_n^{1/2}(\hat{\theta}_{QB} - \theta_0) \overset{D}{\rightarrow} N(0, 1).$$

Observation 1 above is the quasi-posterior extension of the well-known result that, under fairly general conditions, Bayesian point estimates have the same asymptotic distribution as maximum likelihood estimates.

Theorem 1 of Chernozhukov and Hong (2003), which is re-stated here in a slightly different form, is a kind of quasi-Bayesian consistency result, showing that quasi-posterior mass accumulates at the true parameter $\theta_0$.

**Observation 2.** Under the same conditions as Observation 1

$$\|\pi_{M,n}(\theta|y_n) - \pi_{M,\infty}(\theta|y_n)\|_{TV} \overset{p}{\rightarrow} 0,$$

where $\| \cdot \|_{TV}$ indicates the total variation norm, and $\pi_{M,\infty}(\theta|y_n)$ is a normal density with random mean $\theta_0 + Q_n^{-1} \nabla \ell_{M,n}(\theta_0)$ and covariance matrix $Q_n^{-1}$.

Observation 2 may be arrived at informally via a simple Taylor series argument. It is therefore intuitive that the quasi-posterior converges to limiting normal distribution whose covariance matrix is defined by the second derivatives of $\ell_M$. 

The key observation is that the limiting quasi-posterior distribution has a different covariance matrix than the asymptotic sampling distribution of the quasi-Bayes point estimate. The consequence is that the usual Bayesian method of constructing credible intervals based on quantiles of the quasi-posterior sample will, viewed as confidence intervals, not have their nominal frequentist coverage probabilities. Fortunately, thanks to Chernozhukov and Hong (2003), it is known what those two asymptotic covariance matrices look like, which suggests a way to “fix” \( \pi_{M,n}(\theta|y_n) \).

3 The open-faced sandwich adjustment

Let us assume that we have a sample of draws from \( \pi_{M,n}(\theta|y_n) \), generated by replacing the likelihood with \( \ell_M(\theta; y) \) in some MCMC sampler such as Metropolis-Hastings. The aim here is to adjust the quasi-posterior draws such that the adjusted sample realistically reflects how the data informs our uncertainty about the parameter of interest \( \theta \) through the function \( \ell_M(\theta; y) \). Were that the case, the usual credible intervals constructed from empirical quantiles of the adjusted sample would have close to nominal coverage. This will be accomplished by constructing a matrix \( \Omega_n \) that, when applied to the (centered) quasi-posterior sample, will rotate and scale the points in an appropriate way.

We have observed that whereas the asymptotic covariance matrix of \( \hat{\theta}_{M,n} \) is the sandwich matrix \( J_n^{-1} \), the asymptotic covariance matrix of the quasi-posterior distribution is a single “slice of bread” \( Q_n^{-1} \). What we want to do then is complete the sandwich by joining the slice of bread \( Q_n^{-1} \) to the open-faced sandwich \( P_nQ_n^{-1} \) to get \( J_n^{-1} \).

Define \( \Omega_n = Q_n^{-1}P_n^{1/2}Q_n^{1/2} \), the open-faced sandwich adjustment matrix. One can easily check that if \( Z_n \sim N(0, Q_n^{-1}) \), then \( \Omega_nZ_n \sim N(0, J_n^{-1}) \). The idea then is to take samples from \( \pi_M(\theta|y) \) obtained via MCMC and pre-multiply them (after centering) by an estimator \( \hat{\Omega} \) of \( \Omega \) to “correct” the quasi-posterior sample. That is, if \( \theta^{(1)}, \ldots, \theta^{(J)} \) is a sample from \( \pi_M(\theta|y) \), then for each \( j = 1, \ldots, J \),

\[
\theta^{(j)}_{\text{OPFS}} = \hat{\theta}_{QB} + \hat{\Omega}(\theta^{(j)} - \hat{\theta}_{QB})
\]

is the open-face sandwich adjusted sample. It is clear that a consistent estimator of \( \Omega \) will generate credible intervals that are consistent \( (1 - \alpha) \) confidence intervals.
3.1 Estimating $\Omega$

The OFS adjustment (4) requires an estimate of the matrix $\Omega$, which in turn requires estimates of $P$ and $Q$. Because the OFS adjustment occurs post-hoc, it is possible to leverage the existing MCMC sample to compute $\hat{\Omega}$. There are many possible approaches to this task, and here some suggestions are offered, and summarized in Table 1.

While $P$ is notoriously difficult to estimate well (see Kauermann and Carroll 2001, for some examples), Observation 1 immediately suggests a way to estimate $Q$ directly from the MCMC sample with almost no additional computational cost. Specifically, noting that the quasi-posterior density converges to a normal with covariance matrix $Q^{-1}$, a natural estimate $\hat{Q}^{-1}$ is just the sample covariance matrix of the MCMC sample. Another possibility that requires almost no additional computation is to retain the results of the evaluations of $\ell_M$ at each iteration of the sampler and use them to numerically estimate the Hessian matrix at $\hat{\theta}_{QB}$. This Hessian approximation will generally be a good estimator $\hat{Q}_{II}$ of $Q$.

These estimators of $Q$ are not only simple to compute, but they arise as direct results of MCMC output, requiring no additional analytical derivations based on $\ell_M$. They are, in this sense, “model-blind.” Unfortunately, I am unaware of any such “model-blind” estimators of $P$. The simplest solution, in the case where one can write an expression for $\nabla \ell_M(\theta; y)$ and the data $y$ consists of $n$ independent replicates, is to compute a basic moment estimator

$$\hat{P}_I = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_M(\hat{\theta}_{QB}; y_i) \nabla \ell_M(\hat{\theta}_{QB}; y_i)^\prime, \quad (5)$$

which is consistent as $n \to \infty$ under standard regularity conditions. Equation (5) is used in Section 4.1, where there is independent replication. However, because $\nabla \ell_M(\hat{\theta}_{QB}; y)$ converges to zero, when only a single realization of a stochastic process is observed, as in Section 4.1, equation (5) fails to provide a viable estimator. In this latter example, analytical expressions for $P(\theta)$ are available. Plugging $\hat{\theta}_{QB}$ into the analytical asymptotic expression gives an estimator $\hat{P}_{II}$. If a corresponding analytical expression exists for $Q$, the corresponding plug-in estimator is called $\hat{Q}_{III}$.

When an expression for $P(\theta)$ is unavailable, but when it is possible to simulate the process that generated $y$, the parametric bootstrap is an attractive option. Let $y_1, \ldots, y_K$ be $K$ independent
realizations of the stochastic process generated under $\hat{\theta}_{QB}$. Then

$$\hat{P}_{\text{boot}} = \frac{1}{K} \sum_{k=1}^{K} \nabla \ell_M(\hat{\theta}_{QB}; y_k) \nabla \ell_M(\hat{\theta}_{QB}; y_k)'$$

(6)

is the parametric bootstrap estimator of $P$ (an analogous estimator could, of course, be used for $Q$). A nice feature of (5) and (6) is that, at the expense of (perhaps considerable) computational effort, one could substitute finite-difference approximations to the required gradients to obtain reasonable estimators, even in the absence of available closed-form expressions for $\nabla \ell_M(\theta; y)$.

### 3.2 The curvature adjustment

This section describes the curvature-adjusted sampler of Ribatet et al. (2012). This sampler was presented as a way to include composite likelihoods in Bayesian-like models but in fact has far wider generality. Composite likelihoods (Lindsay 1988) are functions of $\theta$ and $y$ constructed as the product of joint densities of subsets of the data. In effect, composite likelihoods treat these subsets as though they were independent. Under fairly general regularity conditions, the asymptotic distribution of maximum composite likelihood estimators have sandwich form (2) (Lindsay 1988). Although Ribatet et al. (2012) consider only composite likelihoods, their argument holds equally well for any function $\ell_M(\theta; y)$ with sandwich asymptotics.

The curvature-adjusted sampler begins by computing the extremum estimator $\hat{\theta}_M$ and $\hat{\Omega}(\hat{\theta}_M)$ as a preliminary step. It works by modifying the Metropolis-Hastings algorithm using a transformation of the form (4) such that the acceptance ratio has a desirable asymptotic distribution. Specifically, at each iteration $j = 1, \ldots, J$, the algorithm proposes a new value $\theta^*$ from some density $q(\cdot | \theta^{(j)})$ and compares it to the current state $\theta^{(j)}$ to evaluate whether to accept or reject the proposal. The difference between the curvature-adjusted sampler and the traditional Metropolis-Hastings sampler is that this comparison now takes place after $\theta^*$ and $\theta^{(j)}$ are scaled and rotated. That is, $\theta^*$ is accepted with probability

$$\min \left\{ 1, \frac{L_M(\theta^*_{CA}; y) \pi(\theta^*) q(\theta^{(j)} | \theta^*)}{L_M(\theta^{(j)}_{CA}; y) \pi(\theta^{(j)}) q(\theta^* | \theta^{(j)})} \right\}$$

(7)

where $\theta^*_{CA} = \hat{\theta}_M + \hat{\Omega}(\hat{\theta}_M)(\theta^* - \hat{\theta}_M)$, and analogously for $\theta^{(j)}_{CA}$. Ribatet et al. (2012) use a result from Kent (1982) to show that the ratio in (7) has the same asymptotic distribution as that of the
true likelihood ratio, and argue that the resultant sample has an asymptotic stationary distribution that is normal with covariance $J^{-1}$, as desired. Note that unlike the OFS adjustment, the curvature-adjusted sampler requires outside initial estimates of $\theta$ and $\Omega$ because the adjustment occurs within the sampling algorithm.

### 3.3 OFS within a Gibbs sampler

With some care, the OFS adjustment may be applied in the Gibbs sampler setting. Suppose that $\theta$ is divided into $B$ blocks such that $\theta_1, \ldots, \theta_B$ forms a partition of $\theta$, and we wish to draw from the quasi-full conditional distribution with density $f(\theta|\theta_{-i}, y) \propto L_M(y|\theta)f(\theta_i)$, where $\theta_{-i}$ refers to the elements of $\theta$ not contained in $\theta_i$. Then the adjustment matrix $\Omega_{\theta_i|\theta_{-i}}$ is defined as before, only now it applies only to $\theta_i$ and is conditional on $\theta_{-i}$. If all full conditional densities $f(\theta|\theta_{-i}, y)$, $i = 1, \ldots, B$, are quasi-full conditional densities in the sense that they are proportional to a product of $\ell_M(y|\theta)$ and another density, the Gibbs sampler may be run by successively drawing from $f(\theta|\theta_{-i}, y)$, $i = 1, \ldots, B$, and the OFS adjustment may proceed post hoc as before. This can be seen by viewing the Gibbs sampler as a special case of Metropolis-Hastings (see Robert and Casella 2004, Section 7.1.4).

Now suppose that at iteration $j$ of a Gibbs sampler we have drawn $\theta_i^{(j)}$ from the quasi-full conditional $f(\theta_i|\theta_{-i}^{(j)}, y)$. Suppose further that $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, y)$ is not a function of $L_M$, as will be the case for many parameters in hierarchical models that contain $L_M$. Since $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, y)$ is a true full conditional density and not a quasi-full conditional density, there is no OFS adjustment to make. But clearly $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, y)$ depends on $\theta_i^{(j)}$, and as a result, plugging in an un-adjusted sample of $\theta_i^{(j)}$ will not result in the desired stationary distribution. It is clear then that to achieve proper variance propagation through the model, we must adjust $\theta_i^{(j)}$ before plugging it into $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, y)$. Therefore, the OFS adjustment within the Gibbs sampler may not be applied post hoc, but rather must occur within the sampling algorithm. Because it has not been proven that this procedure converges to an ergodic chain, care must be taken in practice to assess convergence by examining the output.

Embedding OFS adjustments within Gibbs samplers requires careful consideration of the conditional OFS matrices $\Omega_{\theta_i|\theta_{-i}}, i = 1, \ldots, B$, the adjustment matrices associated with the quasi-full
conditional distributions \( f(\theta | \theta_{-i}) \), \( i = 1, \ldots, B \). Because it is defined conditionally on \( \theta_{-i} \), ideally each \( \Omega_{\theta_i|\theta_{-i}} \) should be re-estimated at each iteration \( j \) based on the current value \( \theta_i^{(j)} \). The matrix \( \Omega_{\theta_i|\theta_{-i}}^{(j)} \) is referred to as the conditional OFS adjustment matrix for \( \theta_i^{(j)} \) at iteration \( j \). Implementing the OFS Gibbs sampler using these conditional OFS adjustments requires estimation of \( \Omega_{\theta_i|\theta_{-i}}^{(j)} \), by one of the techniques described in Section 3.1, for each block of parameters \( i = 1, \ldots, B \) with corresponding quasi-full conditional depending on \( L_M \), and for each Gibbs iteration \( j \) in \( 1, \ldots, J \).

Furthermore, each computation of \( \Omega_{\theta_i|\theta_{-i}}^{(j)} \) requires an estimate of \( \theta_i^{(j)} \) as input, necessitating some sort of optimization or MCMC, again nested within each block \( i = 1, \ldots, B \) and each iteration \( j \) in \( 1, \ldots, J \). This is an enormous computational burden.

Instead, one can make the simplifying assumption that \( \Omega_{\theta_i|\theta_{-i}}^{(j)} \) does not change much from iteration to iteration. Thus one can instead use a constant (with respect to \( j \)) estimate \( \hat{\Omega}_{\theta_i|\theta_{-i}} \), where \( \hat{\theta}_{-i} \) is fixed at its marginal quasi-Bayes estimate. The matrix \( \hat{\Omega}_{\theta_i|\theta_{-i}} \) is referred to as the marginal OFS adjustment matrix for \( \theta_i^{(j)} \).

Despite this simplification, the OFS-adjusted Gibbs sampler still requires additional work relative to an un-adjusted sampler because the algorithm requires \( \hat{\theta}_{QB} \) and \( \hat{\Omega}_{\theta_i|\theta_{-i}} \), \( i = 1, \ldots, B \), as input. In practice, then, the sampler is run twice. The first time, recalling that Observation 2 says the un-adjusted quasi-posterior concentrates its mass at \( \theta_0 \), no OFS adjustments are made, and the generated sample is used to produce \( \hat{\theta}_{QB} \). Next, \( \hat{\theta}_{QB} \) is used to produce \( \hat{\Omega}_{\theta_i|\theta_{-i}} \), \( i = 1, \ldots, B \), using one of the methods described in Section 3.1. Finally, the Gibbs sampler is re-run, this time with the transformation defined by (4) applied for each block \( i \) and each iteration \( j \), \( i = 1, \ldots, B \), \( j = 1, \ldots, J \). Thus, the computational burden required to use marginal OFS adjustments is approximately twice that of the un-adjusted Gibbs sampler. In contrast, the additional computational burden required to use conditional OFS adjustments may range from several fold to several thousand fold, depending on the method used to estimate \( \Omega_{\theta_i|\theta_{-i}}^{(j)} \).

The effects of using the much more computationally efficient marginal adjustment instead of the conditional adjustment have been explored (informally) and found to result in only very minor differences in the resultant adjusted quasi-posteriors. (See Section 4.1 for an example.) The issue of conditional vs. marginal adjustments also appears in Ribatet et al. (2012). They refer to using constant (in \( j \)) adjustment matrices as an “overall” Gibbs sampler and using conditional adjustment
matrices as an “adaptive” Gibbs sampler. Corroborating the findings here, Ribatet et al. (2012), in a more systematic study using a very simple model, found very little difference between their overall and adaptive curvature-adjusted quasi-posteriors.

4 Examples

This section describes two examples of non-likelihood objective functions that have appeared in the literature. In each example, working with the likelihood is problematic for a different reason, and each fits into the OFS framework. In the first example, covariance tapering (Furrer et al. 2006; Kaufman et al. 2008; Shaby and Ruppert 2012) is applied to large spatial datasets. Here the likelihood requires the numerical inversion of a very large matrix. For large datasets, this inversion becomes prohibitively computationally-expensive, so the likelihood is replaced with its tapered version, which leverages sparse-matrix algorithms to speed up computations. In the second example, composite likelihoods for spatial max-stable processes (Smith 1990; Padoan et al. 2010), a probability model is assumed, but the likelihood consists of a combinatorial explosion of terms, and is therefore completely intractable for all but trivial situations. Hence, in this example, the likelihood function is simply not known. These examples may be considered toy models in that, in contrast to the model in Section 5, one could easily maximize their associated objective functions and compute sandwich matrices to obtain point estimates and asymptotic confidence intervals. These examples are used simply to illustrate the effectiveness of the OFS framework.

For each example in this section, a simulation study is conducted to investigate how well the OFS adjustment performs by measuring how often nominal \((1 - \alpha)\) credible intervals cover \(\theta_0\). To do this, datasets \(y_k, k = 1, \ldots, 1000\), are drawn from the model determined by some fixed \(\theta_0\). A random walk Metropolis algorithm is then run, with \(\ell_M(\theta; y_k)\) inserted in place of a likelihood, on each of the 1000 datasets. Next, each set of MCMC samples is used to compute estimates \(\hat{\theta}_{QB,k}\) and \(\hat{\Omega}_k\) using different estimators as discussed in Section 3.1. Finally, each \(\hat{\theta}_{QB,k}\) and \(\hat{\Omega}_k\) is used to adjust their corresponding batch of MCMC output, and the resultant equi-tailed \((1 - \alpha)\) credible intervals are recorded for many values of \(\alpha\). For each example, empirical coverage rates are plotted against nominal coverage probabilities.
To further investigate the quality of the resultant credible intervals, interval scores (Gneiting and Raftery 2007) are computed. For each simulation $i$, the interval score for the $(1 - \alpha)$ interval $(l, u)$ for $j$th component of $\theta_0$ is defined as

$$IS(l_i, u_i; \theta_{0j}) = (u_i - l_i) + \frac{2}{\alpha}(l_i - \theta_{0j})1\{\theta_{0j} < l_i\} + \frac{2}{\alpha}(\theta_{0j} - u_i)1\{\theta_{0j} > u_i\}.$$ (8)

The interval score thus rewards intervals for being narrow (smaller scores are better) and penalizes for failing to cover, with the penalty proportional to the distance between the closer endpoint and the true parameter. For a set of simulations, the average interval score $\frac{1}{1000}\sum_{i=1}^{1000} IS(l_i, u_i; \theta_{0j})/1000$, which is a proper scoring rule (Gneiting and Raftery 2007), is reported.

In addition, the curvature-adjusted sampler of Ribatet et al. (2012) is run for comparison.

4.1 Tapered likelihood for spatial Gaussian processes

The most common structure for modeling spatial association among observations is the Gaussian process (Cressie 1991; Stein 1999). In addition to modeling Gaussian responses, the Gaussian process has been used extensively in hierarchical models to induce spatial correlation for a wide variety of response types (Banerjee et al. 2004).

Here it is assumed that $Y(s) \sim GP(0, C(\theta); s)$, a mean-zero Gaussian process whose second-order stationary covariance is given by a parametric family of functions $C$ indexed by $\theta$, depending on locations $s$ in some spatial domain $D$. It will further be assumed that the covariance between any two observations $y_i$ and $y_j$ located at $s_i$ and $s_j$ is a function of only the distance $||s_i - s_j||$. Then the likelihood for $n$ observations from a single realization of $Y(s)$ is

$$\ell_n(\theta; y_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_n(\theta)|) - \frac{1}{2} y_n'\Sigma_n^{-1}(\theta)y_n,$$ (9)

where $\Sigma_n(\theta) = C(\theta; ||s_i - s_j||)$.

While conceptually simple, these Gaussian process models present computational difficulties when the number of observations of the Gaussian process becomes large, as the likelihood function (9) requires the inversion of a $n \times n$ matrix, which has computational cost $O(n^3)$. To mitigate this...
cost, Kaufman et al. (2008) proposed replacing (9) with the tapered likelihood function

\[ \ell_{tn}(\theta; y_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\Sigma_n(\theta) \circ T_n) \]

\[ - \frac{1}{2} y_n'((\Sigma_n(\theta) \circ T_n)^{-1} \circ T_n)y_n, \]  

where the \( \circ \) notation denotes the element-wise product, and \( T_{ij} = \rho_t(||s_i - s_j||) \), a compactly-supported correlation function that takes a non-zero value when \( ||s_i - s_j|| \) is less than some pre-specified “taper range.” The compact support of \( \rho_t \) induces sparsity in \( T_n \), and hence all operations required to compute (10) may be computed using specialized sparse-matrix algorithms, which are much faster and more memory-efficient than their dense-matrix analogues.

Under suitable conditions, the tapered likelihood satisfies asymptotics of the form (2), and Observations 1 and 2 apply (Shaby and Ruppert 2012). For the simulations, we take \( C(\theta; ||s_i - s_j||) = \sigma^2 \exp(-c/\sigma^2 \cdot ||s_i - s_j||) \), with \( \theta = (\sigma^2, c)' = (1, 0.2)' \). The observations are made on a 40 \( \times \) 40 unit grid, so that each dataset \( y \) is a single 1600-dimensional realization of a stochastic process. Weakly informative half-Cauchy priors (see Gelman 2006) were used for both parameters.

For this example, analytical expressions for both \( P(\theta) \) and \( Q(\theta) \) are available (Shaby and Ruppert 2012). As described in Section 3.1, the plug-in estimator \( \hat{\Omega}_k = \hat{Q}_{III}(\hat{\theta}_{QB,k})^{-1}\hat{P}_{II}(\hat{\theta}_{QB,k})^{1/2}\hat{Q}_{III}(\hat{\theta}_{QB,k})^{1/2} \), is used, as well as \( \hat{\Omega}_k = \hat{Q}_I^{-1}\hat{P}_{II}(\hat{\theta}_{QB,k})^{1/2}\hat{Q}_I^{-1} \), with \( \hat{Q}_I^{-1} \) computed directly from the MCMC sample, for each \( k = 1, \ldots, 1000 \) simulated datasets.

Figure 1 shows that the un-adjusted MCMC samples (dotted curves) yield horrible coverage properties for both \( \sigma^2 \) and \( c \). It is somewhat interesting that while the “naive” intervals severely under-cover \( \sigma^2 \), they severely over-cover \( c \). We therefore see that a naive implementation results in being overly optimistic about estimates of \( \sigma^2 \) while being overly pessimistic about estimates of \( c \). The OFS-adjusted intervals display much more accurate coverage, achieving nearly nominal rates, although for \( c \), the asymptotic expression for \( \hat{Q} \) seems to produce intervals that are systematically slightly too short. The curvature-adjusted sampler results in similar coverage.

Table 2 shows that 95% intervals based on the an-adjusted sampler score badly, and intervals based on the OFS and curvature adjustments perform comparably. Together with Figure 1, Table 2 indicates that for this example the OFS and curvature adjustments result in accurate intervals of similar quality according to the interval score.
To explore how the marginal OFS adjustment differs from the conditional adjustment in the Gibbs sampler setting, data was simulated from a spatial linear model, \( Y(s) \sim \text{GP}(X\beta, C(\theta); s) \). The regression coefficients were set at \( \beta = (-0.5, 0, 0.5)' \), and the same spatial design and covariance function as above was used. The design matrix \( X \) was a \( 1600 \times 3 \) matrix of standard normal deviates, and the prior distribution for \( \beta \) was a vague normal centered at zero. At each Gibbs iteration, \( \theta \) was updated using a Metropolis step using the tapered likelihood, and \( \beta \) was updated by drawing directly from its conditionally conjugate full conditional distribution. The sampler was first run without adjustment, and \( \hat{\theta}_{QB} \) and \( \hat{\beta}_{QB} \) were computed as the quasi-posterior means. The marginal OFS adjustment matrix was then computed using the analytical expression from Shaby and Ruppert (2012) by plugging in \( \hat{\theta}_{QB} \) and \( \hat{\beta}_{QB} \). Next, the Gibbs sampler was run a second time using the estimated marginal OFS matrix to adjust the sample from the full conditional distribution of \( \theta \) at each iteration. Finally, the Gibbs sampler was run a third time, this time estimating the conditional OFS adjustment matrix at each iteration by maximizing the tapered likelihood function and plugging the resultant parameter estimates into the asymptotic formula for \( \hat{\Omega} \).

Because the conditional OFS-adjusted Gibbs sampler is so computationally expensive, just a few datasets were simulated, and the output from one of them is reported. Figures 2(a) and 2(b) compare the marginal adjusted quasi-posterior distributions for the two covariance parameters, generated with the marginal and conditional OFS adjustments. The qq-plot for \( \sigma^2 \) shows almost perfect agreement except for a handful of MCMC samples on the upper tail. For the \( c \) parameter, the qq-plot shows that the marginal OFS adjustment produces a quasi-posterior that is the same shape as that of the conditional adjustment, but is slightly shifted to the right. Figure 2(c) shows contours of a kernel density estimate of the joint marginal OFS-adjusted quasi-posterior distribution of the two covariance parameters, with the marginal adjustment in black and the conditional adjustment in gray. The contours are very similar, with some small differences appearing on the right half of the \( \sigma^2 \)-axis, indicating good agreement between the two bivariate distributions. The output from all the simulated datasets looked qualitatively similar, with no noticeable systematic differences between the two adjustments. The choice of adjustment had no discernible effect on the quasi-posterior distribution of \( \beta \).
4.2 Composite likelihood for max-stable processes

Statistical models for extreme values that include spatial dependence are useful for studying, for example, extreme weather events like heat waves and powerful storms (Cooley et al. 2007; Sang and Gelfand 2010, e.g.). Extreme value theory says that the distribution of block-wise maximum values (such as annual high temperatures) of independent draws from any distribution converges to a generalized extreme value (GEV) distribution (see Coles 2001), if it converges at all. The asymptotics therefore suggest that any model of block-wise maxima at several spatial locations ought to have GEV marginal distributions with distribution function

\[ F(y; \mu, \sigma, \xi) = \exp \left\{ - \left[ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \]

where \( \mu \) is a location, \( \sigma \) a scale, and \( \xi \) a shape parameter that determines the thickness of the right tail. The GEV may be characterized by the max-stability property (Coles 2001). More generally, block-wise maxima of random vectors must also converge to max-stable distributions.

Sang and Gelfand (2010) achieve spatial dependence with GEV marginals through a Gaussian copula construction. However, Gaussian copula models have been strongly criticized (Klüppelberg and Rootzén 2006) because they do not result in max-stable finite-dimensional distributions, nor do they permit dependence in the most extreme values, referred to as tail dependence, and it is clear that physical phenomena of interest do exhibit strong spatial dependence even among the most extreme events.

An alternate approach for encoding spatial dependence of extreme values is through max-stable process models (de Haan 1984), which are stochastic processes over some index set where all finite-dimensional distributions are max-stable. Explicit specifications of spatial max-stable processes based on the de Haan (1984) spectral representation have been proposed by Smith (1990), Schlather (2002), and Kabluchko et al. (2009). These formulations have the advantage that they do represent tail dependence.

Unfortunately, for all of the available spatial max-stable process models, joint density functions of observations at three or more spatial locations are not known (a slight exception is the Gaussian extreme value process (GEVP) (Smith 1990), for which Genton et al. (2011) derives trivariate densities). Since bivariate densities are known, Padoan et al. (2010) proposes parameter estimation
and inference via the pairwise likelihood, where all bivariate log likelihoods are summed as though they were independent:

\[ \ell_p(\theta; y) = \sum_{i \neq j} \ell(y_i, y_j; \theta). \]

The pairwise likelihood is a special case of a composite likelihood (Lindsay 1988). Padoan et al. (2010) show that asymptotic normality of the form (2) applies, so we may again apply the OFS adjustment.

The following simulation experiment consists of 1000 draws from a GEVP with unit Fréchet margins on a 10 × 10 square grid, with 100 replicates per draw. An example of a single replicate is shown in Figure S-6 in the Supplementary Materials. This setup would correspond, for example, to 100 years of annual maximum temperature data from 100 weather stations.

The unknown parameter \( \theta \) in a 2-dimensional GEVP is a 2 × 2 covariance matrix. For this simulation, \( \theta_0 = (\Sigma_{11}, \Sigma_{12}, \Sigma_{22})' = (0.75, -0.5, 1.25)' \). The prior distribution for \( \Sigma \) is a vague inverse-Wishart. For each draw, a long MCMC chain is run and \( \hat{\theta} \) is computed as the posterior mean. In addition, for each draw, all four \( Q - P \) combinations of \( \hat{Q}_I, \hat{Q}_II, \hat{P}_I, \) and \( \hat{P}_{\text{boot}} \), as defined above, are computed to produce four estimates of \( \Omega \). Finally, the curvature-adjusted MCMC sampler from Ribatet et al. (2012) is run on each simulated dataset, with \( \Omega \) estimated from \( \hat{Q}_II \) and \( \hat{P}_I \), evaluated at the maximum pairwise likelihood estimate of \( \Sigma \).

Figure 3 shows that, for this simulation, the OFS adjustment produces credible intervals that cover at almost exactly their nominal rates. Furthermore, OFS-adjusted credible intervals based on the four values of \( \hat{\Omega} \) turned out nearly identical. The curvature-adjusted sampler also achieves nominal coverage. The un-adjusted intervals systematically under-cover for each of the three parameters. This is expected (as noted by Ribatet et al. (2012)), as the pairwise likelihood over-uses the data by including each location in roughly \( n/2 \) terms of the objective function rather than just one, as would be the case with a likelihood function. This results in a pairwise likelihood surface that is far too sharply-peaked relative to a likelihood surface. The OFS adjustment seems to successfully compensate for this effect.

Interval scores for the GEVP are shown in Table 3. Just as in Section 4.1, the un-adjusted 95% intervals score poorly. However, unlike in the tapering example, Table 3 shows marked differences between the OFS and curvature adjustments. Given that both methods yield almost exactly nominal
coverage, the higher interval scores reveal that for this example, the curvature-adjustment produces intervals that are wider than those of the OFS adjustment and, when they fail to cover, miss by a greater margin.

To investigate how the performance of the OFS adjustment varies with $n$, the same experiment was repeated with 25 replications and again with 50 replications, rather than the 100 replications reported here. The results from the smaller sample sizes are consistent with those reported here, with all OFS adjustments and the curvature adjustment producing close to nominal coverage, but with curvature adjusted intervals having comparatively worse interval scores. These results are reported in the Supplementary Materials.

5 Data analysis

Next, tapered quasi-Bayesian analysis is applied to a hierarchical model that includes a Gaussian process. Instead of a purely spatial random field as in Section 4.1, a spatio-temporal random field is assumed, which highlights some of the advantages of tapering over other methods designed for large spatial datasets.

The dataset comes from a “citizen science” initiative called eBird (www.ebird.org). The idea of citizen science is that many non-professional observers can be leveraged to collect an enormous amount of data. eBird participants across North America record the birds they see, along with the time and location of the observation, into a web-based database. Here, 6114 observations of the Northern Cardinal in a section of the eastern United States over a period from 2004 to 2007 (Figure S-3 in the Supplementary Materials) are analyzed.

Inspection of the data suggests an overdispersed Poisson model. Let $Y_1, \ldots, Y_n$ be observed counts and $X = [x_1, \ldots, x_n]$ be a matrix of covariates associated with each observation. Also, let $s = [s_1, \ldots, s_n]$ and $t = [t_1, \ldots, t_n]$ be the spatial and temporal locations, respectively, associated with $Y_1, \ldots, Y_n$, with space indexed by latitude and longitude.

For this example, a small number of predictors are deliberately chosen, several of which vary spatially. Preliminary analyses led to a set of 10 covariates that includes time of day, day of year, human population density, percentage of developed open space (single-family houses, parks, golf
courses, etc.), tree canopy density, and variables that measure observer effort. Simple transformations (logs, powers, etc.) were applied to some of the covariates, as suggested by ornithologists and exploratory analyses.

The model is specified as

\[ y_i | \lambda_i \sim \text{Pois}(\lambda_i) \]

\[ \log \lambda_i(x_i, s_i, t_i) = x_i' \beta + Z_i(s_i, t_i) + \varepsilon_i \]

\[ z(s, t) | \theta^* \sim N(0, \Sigma^*(\theta^*; s, t)) \]

\[ \varepsilon_i | \tau \sim \text{iid } N(0, \sigma^2_\varepsilon). \]

It is assumed that the random effect \( z(s, t) \) has a Gaussian random field structure. Thus, this model is an example of “model-based geostatistics” of Diggle et al. (1998), a class of spatial generalized linear models that has seen wide application in the environmental literature. Again, extremum estimators like maximum likelihood are not possible for this model because the random effect \( z(s, t) \) cannot be marginalized out analytically or numerically.

Even though Northern Cardinals are not migratory birds, a spatio-temporal structure for the random effect has a great intuitive appeal. One can easily imagine clusters of birds habitating different locales, moving from place to place based on things like food availability or disturbances. The correlation range of the spatio-temporal random effect informs the ornithologists about the scales of movements of Northern Cardinals in space and time, as well as providing clues about what un-measured covariates are needed to explain the pattern of Northern Cardinal observations.

The parameter \( \varepsilon \) can be interpreted as either an overdispersion parameter, or as the traditional “nugget” effect, representing small-scale variation or measurement error. It will be convenient to do some marginalization and consider the distribution of the log-means directly. Furthermore, the matrix \( \Sigma^*(\theta^*; s, t) + \sigma^2_\varepsilon I \) can be written simply as \( \Sigma(\theta; s, t) \) and \( \theta^* \) and \( \sigma^2_\varepsilon \) condensed into the single parameter vector \( \theta \). The resulting model, equivalent to (11), is written as

\[ y_i | \lambda_i \sim \text{Pois}(\lambda_i) \]

\[ \log \lambda_i(x_i, s_i, t_i) \equiv b_i \]

\[ b(X, s, t) | \theta, \beta \sim N(X\beta, \Sigma(\theta; s, t)). \]
Another level in the hierarchy imposes a ridge penalty on the regression coefficients $\beta$, specified as $\beta \sim N(0, \sigma_\beta^2 I)$. Finally, priors are needed for the parameters $\theta$ and $\sigma_\beta^2$.

For $\Sigma(\theta; s, t)$, a spatio-temporal covariance model is chosen from Gneiting (2002). The covariance functions described therein are nonseparable in that (except in special cases) they cannot be written as the product of a purely spatial and purely temporal covariance function. Specifically, let

$$C(\theta^*; h, u) = \frac{\sigma^2}{(au^{2\alpha} + 1)^2} \cdot \exp \left\{ -\frac{c}{\sigma^2} h^{2\gamma} \right\},$$

where $h$ and $u$ are distances between observation points in space and time, respectively. The parameters $\alpha \in (0, 1]$ and $\gamma \in (0, 1]$ control the smoothness of the process. These parameters were at convenient values (as suggested by Gneiting (2002)) of 1 and .5, respectively, because they were not well-identified by the data.

The parameter $\omega \in [0, 1]$ has the nice interpretation of specifying the degree of nonseparability between purely spatial and purely temporal components; when $\omega = 0$, $C(\theta; h, u)$ is the product of a purely temporal and a purely spatial (exponential) covariance function.

Priors for the parameters $\sigma^2$, $a$, $c$, $\sigma_\epsilon^2$, and $\sigma_\beta^2$ are specified as vague Cauchy distributions, truncated to have only positive support. The interaction parameter $\omega$ is given a uniform prior on $[0, 1]$. A valid spatio-temporal taper matrix may be constructed as the element-wise product of a spatial and a temporal taper matrix $T = T_s \circ T_t$. Constructed this way, $T$ inherits the sparse entries of both $T_s$ and $T_t$, and may therefore itself be extremely sparse.

Several other methods exist to mitigate the computational burden imposed by large spatial datasets with non-Gaussian responses. Wikle (2002) and Royle and Wikle (2005) embed a continuous spatial process into a latent grid and work in the spectral domain using fast Fourier methods. However, applying Fourier methods here is problematic, as it is not obvious how to do so for a process that has spatio-temporal structures. Low rank methods like predictive processes (Banerjee et al. 2008; Finley et al. 2009) and fixed rank Kriging (Cressie and Johannesson 2008) are also popular for spatial data. Applying these methods to spatio-temporal models is possible, but awkward. For predictive processes, one must decide how to specify knot locations in space $\times$ time. For fixed rank Kriging, one must specify knot locations as well as space-time kernel functions. Fixed rank Kriging has been adapted to the spatio-temporal setting (Cressie et al. 2010) through a linear filtering framework, but only for Gaussian responses. Finally, Gauss-Markov approximations to
continuous spatial processes are fast to compute, especially when using Laplace approximations in place of MCMC (Lindgren et al. 2011; Rue et al. 2009). However, again, these methods do not apply to data with spatio-temporal random effects.

In contrast, application of the tapering approach in the spatio-temporal context is immediate and even potentially enjoys increased computational efficiency relative to the purely spatial context because of the additional sparsity induced by element-wise multiplication with the temporal taper matrix.

For the eBird data, a taper range of 20 miles and 60 days gives a tapered covariance matrix with about .5% nonzero elements. The tapered likelihood (10) produces un-biased estimating functions regardless of taper range, so choosing relatively short taper ranges does not limit the correlation length that the model can estimate. MCMC was carried out using a block Gibbs sampler. Each evaluation of the expensive normal log likelihood was replaced by its tapered analogue. Within each Gibbs iteration, each of \( b, \theta \), and \( \sigma^2 \) were updated with a random walk Metropolis step. The full conditional distribution for \( \beta \) is conditionally conjugate, enabling a simple update as a draw from the appropriate normal distribution.

As described in Section 3.3, the tapered Gibbs sampler was run twice. Samples from the first run were used to produce point estimates of \( \theta \) and the marginal OFS adjustment matrix \( \Omega_{\theta|\beta,b,\sigma^2_\beta} \). The estimate \( \hat{\Omega}_{\theta|\beta,b,\sigma^2_\beta} \) was computed from the asymptotic expressions for \( Q \) and \( Q \) evaluated at \( \hat{\theta}_{QB} \), the quasi-posterior mean. Because the quasi-posterior distribution of the interaction parameter \( \omega \) was nearly uniform on [0,1], it was excluded from the adjustment. The conditional OFS was not attempted, as doing so would have required several months of computation time.

After discarding 5000 burn-in iterations, 5000 MCMC samples were used for estimation and prediction. Model fit was assessed by setting aside 1000 randomly chosen observations as a test set and making draws of the posterior predictive distribution at the holdout locations. To evaluate the results, three quantities were calculated. The first is the empirical coverage of nominal 95% predictive intervals. The second is the interval score for 95% predictive intervals. The third is the percent deviance explained for the posterior predictive mean relative to the null model, defined as \( [D(y;\hat{y}) - D(y;\tilde{y})]/D(y;\tilde{y}) \), where \( D(y;\cdot) \) is the Poisson deviance function, \( \tilde{y} \) is the posterior predictive mean, and \( \hat{y} \) is the sample mean of the training set, representing prediction based on
the null model. Percent deviance explained is a way of measuring predictive accuracy when the response is non-Gaussian. Results are given for both the un-adjusted sampler and the OFS-adjusted sampler, and are summarized in Table 4.

Table 4 shows that the model fits well, and that the OFS adjustment makes almost no difference in the quality of the resultant predictive distribution. The lack of impact on the test set predictions is not altogether surprising, as the adjustment is made only to the conditional quasi-posterior distribution of the covariance parameters of the random effect, and small perturbations of covariance parameters of Gaussian processes have little effect on spatial predictions (Kaufman and Shaby 2013). However, because the covariance parameters have interpretive interest in this application, assessing their uncertainty accurately is a key component of the analysis, regardless of the lack of impact on the predictive surface.

The MCMC was then re-run using the entire dataset. Pointwise quantiles of the posterior correlation surface are shown in Figure 4, for both the un-adjusted and adjusted samples. The point at which the correlation drops to .05, often called the “effective range” of a process, is the most extreme contour displayed in each of the plots in Figure 4. While the two sets of contours do not differ much in the median, they are quite different in the upper and lower quantiles. For this analysis, the correlation structure is a key component with a useful interpretation, so its posterior uncertainty is of interest. Comparing the OFS-adjusted and un-adjusted correlation surfaces (Figure 4), it is interesting to note that OFS adjustment reveals decreased uncertainty relative to the un-adjusted sample in the temporal dimension of the correlation surface, but increased uncertainty in spatial dimension.

The fairly long median effective range of around 225 days at spatial lag 0 seemed reasonable to a panel of ornithologists, as Northern Cardinals, while they do move around to some degree, are not migratory birds. The effective range of 3 miles at time lag 0 seemed reasonable as well. Northern Cardinals build new nests each year and are socially monogamous within a breeding season, but divorces sometimes occur between years. They generally stay close to the nest to forage and bathe. Males are highly territorial and will occasionally challenge neighboring males’ breeding territories. These behaviors are consistent with a posterior median temporal dependence range of a significant fraction of a single year, and a posterior median spatial range that is larger
than but in the ballpark of an individual’s territorial range.

A discussion of the posterior estimates for some of the more interesting fixed effects, as well as a discussion of posterior predicted abundance maps, may be found in the Supplementary Materials.

6 Discussion

The open-faced sandwich adjustment provides a way to incorporate estimating functions that are not likelihoods into Bayesian-like models. While the resulting inference does not enjoy the elegant formal probabilistic interpretation of pure Bayesian analysis, it does inherit some of its most desirable attributes: borrowing strength across parameters, the ability to work with complicated hierarchical structures, and propagating uncertainty throughout model components, to name a few. When the likelihood function is unknown or has undesirable properties, the OFS adjustment allows one to retain these beneficial features of Bayesian analysis while avoiding the need to compute the likelihood function by substituting a suitable objective function in its place.

These benefits come with a few costs. First, the resultant MCMC samples may not be interpreted as though they came from a true Bayesian model. Second, the coverage characteristics of the output are only as good as the applicability of the asymptotic approximation and the practitioner’s ability to estimate the sandwich matrix, which can be a difficult task in some situations (Kauermann and Carroll 2001). Third, estimating the adjustment matrix using sample moments or a bootstrap and relying on it to produce posterior samples has a decidedly “un-Bayesian” feel to it. Finally, using the adjustment in the Gibbs sampler context does require approximately twice the computational effort as the un-adjusted Gibbs sampler, which in some cases can be considerable.

In addition to these considerations, comparisons between the OFS adjustment and the curvature adjustment of Ribatet et al. (2012) seem natural. In the simulations presented here, both adjustments performed extremely well. Both covered at almost exactly nominal levels, but OFS produced superior intervals in the max-stable process example, as measured by the interval score. The curvature adjustment shares with OFS both the advantages and disadvantages described above. But in addition, the curvature adjustment, as implemented in the data example in Ribatet et al. (2012), has several additional drawbacks to consider. First, Ribatet et al. (2012) require an outside method
to estimate of $\theta$ and $\Omega$, whereas the OFS adjustment uses the MCMC sample to estimate $\theta$ and $\Omega$. Using the un-adjusted quasi-posterior sample to estimate $\theta$ and $\Omega$ as has been done here takes advantage of borrowing strength, leveraging prior information, etc. that simply maximizing $\ell_M(\theta)$ cannot. It should be noted, however, that this drawback in the curvature adjustment can be avoided. One could easily apply the strategy suggested here in Section 3.3, running the sampler first without adjustment to estimate $\theta$ and $\Omega$ in a Bayesian-like way, and then using these estimates to implement the curvature-adjusted sampler. In the Metropolis context, this strategy requires twice the computational effort of OFS; since OFS is applied to the sample post hoc, there is no need to run the sampler a second time. In the Gibbs sampler setting, however, the computational burden is identical.

The most obvious drawback of the curvature adjustment is the enormous computational cost imposed by estimating the conditional adjustment matrix in the adaptive version of the curvature adjusted Gibbs sampler favored by Ribatet et al. (2012). This simply would not have been feasible, for example, in the eBird example of Section 5. This complication can probably be avoided by using the marginal, rather than the conditional, version of $\hat{\Omega}$. In fact, since their simulated comparisons between the marginal and conditional forms of $\hat{\Omega}$ performed so similarly, it is unclear why Ribatet et al. (2012) use the much more computationally expensive conditional version in their data example. In the end, conditional on implementation details, the OFS and curvature adjustments are quite similar both in performance and in spirit.

One problem that remains is a theoretical treatment of the convergence properties of the OFS-adjusted Gibbs sampler. Beyond the usual MCMC diagnostic tools, there are currently no guarantees about what distribution the chain converges to, or indeed whether it converges at all. This is a topic for future research.

**Acknowledgements**

The author wishes to thank Cari Kaufman, Richard Smith, Dan Cooley, and Alan Gelfand for much helpful discussion. Patient guidance on the eBird analysis was provided by Wesley Hochachka, Steve Kelling, and especially Daniel Fink at the Cornell Lab of Ornithology. Implementation of
the GEVP was aided by generous assistance from Mathieu Ribatet. This manuscript was improved by the helpful comments of three anonymous reviewers.

Supplementary Materials

Appendix: Additional figures, tables, and analyses that were edited out of the main text due to length considerations. (.pdf)

R-files: R scripts to re-produce the simulations and analyses contained in the text. (.tar.gz)

References


<table>
<thead>
<tr>
<th>Estimator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Q}^I$</td>
<td>sample covariance of MCMC sample</td>
</tr>
<tr>
<td>$\hat{Q}^I$</td>
<td>Hessian of $\ell_M(\hat{\theta}_{QB})$</td>
</tr>
<tr>
<td>$\hat{Q}^I$</td>
<td>plug $\hat{\theta}_{QB}$ into asymptotic formula</td>
</tr>
<tr>
<td>$\hat{P}^I$</td>
<td>moment estimator based on score vector</td>
</tr>
<tr>
<td>$\hat{P}^I$</td>
<td>plug $\hat{\theta}_{QB}$ into asymptotic formula</td>
</tr>
<tr>
<td>$\hat{P}_{\text{boot}}$</td>
<td>parametric bootstrap</td>
</tr>
</tbody>
</table>

Table 1: Summary of estimators of sandwich components.

<table>
<thead>
<tr>
<th></th>
<th>un-adj</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>curv. adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
<td>3.62</td>
<td>1.05</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>$c$</td>
<td>0.15</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 2: Interval scores for equi-tailed 95% credible intervals for $\sigma^2$ and $c$ constructed using un-adjusted MCMC samples, OFS-adjusted samples, and curvature-adjusted samples.

<table>
<thead>
<tr>
<th></th>
<th>un-adj</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>$\hat{Q}^I, \hat{P}^I$</th>
<th>curv. adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma_{11}$</td>
<td>0.90</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.37</td>
</tr>
<tr>
<td>$\Sigma_{12}$</td>
<td>0.82</td>
<td>0.23</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.34</td>
</tr>
<tr>
<td>$\Sigma_{22}$</td>
<td>1.47</td>
<td>0.36</td>
<td>0.36</td>
<td>0.35</td>
<td>0.35</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 3: Interval scores for equi-tailed 95% credible intervals for the GEVP parameters, constructed using un-adjusted MCMC samples, OFS-adjusted samples, and curvature-adjusted samples.
Table 4: Empirical coverage and interval scores for equi-tailed 95% posterior predictive intervals, as well as percent deviance explained, for the holdout set, constructed using un-adjusted and OFS-adjusted MCMC predictive samples.

<table>
<thead>
<tr>
<th></th>
<th>cov. 95%</th>
<th>IS 95%</th>
<th>%Dev. expl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>un-adj</td>
<td>95.8</td>
<td>15.65</td>
<td>61.26</td>
</tr>
<tr>
<td>OFS</td>
<td>95.4</td>
<td>15.97</td>
<td>60.58</td>
</tr>
</tbody>
</table>

Figure 1: Empirical coverage rates for equi-tailed credible intervals based on MCMC samples using the tapered likelihood. Blue and red curves are OFS-adjusted samples using different estimates of $\Omega$, green curves are from a curvature-adjusted sampler, and dotted curves are un-adjusted samples.
Figure 2: Comparison of marginal and conditional OFS-adjusted quasi-posterior distributions. Panels (a) and (b) show qq-plots of the marginal quasi-posteriors of the two covariance parameters. Panel (c) shows contour plots of a kernel density estimate of the joint quasi-posterior of the same parameters.
Figure 3: Empirical coverage rates for equi-tailed credible intervals based on MCMC samples using the pairwise likelihood. Colored curves are OFS-adjusted samples using different estimates of $\Omega$, and from a curvature-adjusted sampler. Dotted curves are un-adjusted samples.
Figure 4: Posterior quantiles of spatio-temporal correlation surface. The top row was computed from the un-adjusted sampler, and the bottom row was computed from the OFS-adjusted sampler. The median surfaces are similar, but the high and low quantile contours show that the uncertainty is noticeably altered by the adjustment.