Technical Vignette 5: Understanding intrinsic Gaussian Markov random field spatial models, including intrinsic conditional autoregressive models

Christopher Paciorek, Department of Statistics, University of California, Berkeley, and Department of Biostatistics, Harvard School of Public Health

Version 1.0: December 2009

Spatial models for areal data, such as disease mapping based on aggregated disease counts in administrative districts, commonly employ Markov random field (MRF) models, in particular the conditional specification of a Markov random field known as the conditional autoregressive (CAR) model. While CAR and MRF refer to the same model structure, one often thinks of CAR models in terms of the conditional distributions for each of the random variables given the random variables in neighboring areas. A comprehensive reference for Gaussian MRFs (GMRFs) is Rue and Held (2005). The text below also draws from Banerjee, Carlin, and Gelfand (2004, Sections 3.2, 3.3, 5.4). One of the major appeals of the MRF specification is that the precision matrix for the random field is very sparse, which enables efficient computation through well-developed sparse matrix routines, such as in the spam package in R.

Here I highlight some of the features of GMRFs in spatial modeling, in particular technical aspects related to specification of intrinsic GMRFs, which are improper models. I've found that impropriety in the intrinsic GMRF can be hard to grasp and I hope to lay it out clearly and relatively briefly here in one place.

Basic spatial model structure (standard CAR model)

A basic GMRF model for a spatial collection of random variables used to model areal data, with each random variable representing one of n spatial areas that partition a given domain, $\theta = \{\theta(s_1), \ldots, \theta(s_n)\}$, can be specified in terms of the (scaled) precision, Q, $\theta \sim \mathcal{N}_n(0, \tau^2 Q^{-1})$. Here I use the inverse of Q conceptually because in most cases Q is of less than full rank, discussed below. Non-zero mean, including the effect of covariates can be specified separately as part of the larger model as needed. A standard form for a MRF model for areal data is to specify that any pair of elements of θ, say θ_i and θ_j, be independent given the remaining elements, if the areas represented by the elements do not share a spatial border. This conditional independence corresponds to the ijth element of Q being zero. Then one usually takes a weight of one for each pair that share a common border. The resulting Q has $Q_{ij} = -1$ for areas that share a common border (or are considered neighbors under some other criterion) and Q_{ii} equal to the number of neighbors for area i, m_i. The resulting full conditionals specify the standard CAR model:

$$\theta_i|\theta_{-i} \sim \mathcal{N}(\sum_{j \in N(i)} \theta_j/m_i, \tau^2/m_i)$$

where $N(i)$ is the set of neighbors of area i. Note that Q can be expressed as $Q = D^{-1}(I - B) = D^{-1} - W$ where D is a diagonal matrix with elements $1/m_i$ and B is related to the proximity (or adjacency) matrix, W, by $B_{ij} = W_{ij}/W_{i+}$where W_{i+} is the row sum for the ith row. W is simply a matrix of all zeroes, except ones for W_{ij} where areas i and j share a border, so $W_{i+} = m_i$. The diagonal of W is all zeroes. More generally, whether two areas are 'neighbors' can be determined by another rule, such as based on distance and the weights in W do not have to be ones; these are subjective user choices.

MRF approximation to a thin plate spline

Rue and Held (2005) and Yue and Speckman (2008) describe a more flexible MRF model that approximate a thin plate spline, specified on a regular grid. Rather than using a proximity matrix with ones only for nearest, cardinal neighbors, the proximity matrix, W, has the value 8 for pairs that are nearest (first order) neighbors in the cardinal directions, the value -2 for pairs that are nearest neighbors on the diagonal and the value -1 for pairs that are second order neighbors in the cardinal directions. For this model the generalization of the number of neighbors is $W_{i+} = D_{ii}^{-1} = Q_{ii} = 20$. Boundary corrections are needed for grid cells on the boundary or one cell away from the boundary. For such interior cells the model's full conditionals are such that the conditional mean of θ_i given the neighbors above is $8/20$ times the nearest cardinal neighbors, minus $2/20$ times the nearest diagonal neighbors, minus $1/20$ times the second nearest cardinal neighbors, with conditional variance $\tau^2/20$. Rue and Held (2005, p. 114) describe the derivation of this model based on the forward difference analogue of penalizing the derivatives of a surface to derive the thin plate spline.

Note that this approach generalizes a second order random walk model in one dimension. In one dimension, this gives us a model specified in time order as $\theta_t|\theta_{\text{past}} \sim \mathcal{N}(2\theta_{t-1} - \theta_{t-2}, \tau^2)$ or based on the full conditionals as $\theta_t|\theta_{-t} \sim \mathcal{N}(\frac{4}{6}(\theta_{t+1} + \theta_{t-1}) - \frac{1}{6}(\theta_{t+2} + \theta_{t-2}), \tau^2/6)$ for times at least two away from the first or last time and necessary boundary corrections otherwise (see Rue and Held (2005, p. 110)). The elements of the tth row of Q (for $t > 1$ and $t < n - 1$) are zeroes except $Q_{t,t-2} = Q_{t,t+2} = 1$, $Q_{t,t-1} = Q_{t,t+1} = -4$, and $Q_{tt} = 6$. This model is also given in the Ice example in the BUGS manual.

Impropriety

The matrix Q is of less than full rank. For the simple CAR model one eigenvalue of the precision matrix is zero, corresponding to the linear combination, $\sum \theta_i$. For the simple CAR model, the model can be seen as a joint distribution, $P(\theta) \propto \exp(-\frac{1}{2\tau^2}\sum_{i \neq j} w_{ij}(\theta_i - \theta_j)^2)$, which is invariant to the addition of a constant to all the θs ($Q1 = 0$), and therefore improper, not giving a finite variance for the linear combination, $\sum \theta_i$. For the thin plate spline MRF, there are three zero eigenvalues, corresponding to the sum and to linear terms in the cardinal directions, with the prior being invariant to addition of a plane to the θs. This means that the prior does not constrain these linear combinations of the θs, with infinite variance for these linear combinations. The solution to this issue is to use the pseudoinverse when one needs to compute solutions to linear systems of the form, $Q^{-1}x$. In this setting, the following generalized inverse satisfies the Moore-Penrose conditions and is the unique pseudoinverse. First, compute the eigendecomposition, $Q = \Gamma\Lambda\Gamma^T$. Then the pseudoinverse is $"Q^{-1}" = Q^+ = \Gamma\Lambda^+\Gamma^T$ where Λ^+ is a diagonal matrix with $\lambda_i^+ = 1/\lambda_i$ for $\lambda_i \neq 0$ and $\lambda_i^+ = 0$ when $\lambda_i = 0$. Numerically, one takes the diagonal element to be zero whenever λ_i is within some tolerance of zero. What are we doing statistically? We are forcing the appropriate linear combinations to have zero variance, thereby giving us a proper distribution on the reduced dimension subspace. To allow flexibility in modeling, we need only add additional parameters to the mean model for θ to take the place of the omitted linear combinations. This gives us a proper prior without sacrificing flexibility. The implicit model using the pseudoinverse is $\theta \sim \mathcal{N}_{n-c}(0, \tau^2 Q^+)$, where c is the number of constraints (the number of zero eigenvalues of Q).

What about the normalizing constant in this context? When we enforce the constraint(s), we make use of Λ^+, so $|Q^+| = |\Gamma\Lambda^+\Gamma| \equiv \prod_{i=1}^{n-c} \lambda_i^+$, which makes sense because we are working in a reduced dimension subspace and ignore the eigenvalues corresponding to the constrained portion of the space. Accordingly, the appropriate normalizing constant for these specifications involves $(\tau^2)^{-(n-c)/2}$ rather than $(\tau^2)^{n/2}$, since the prior is proper in the subspace. Of course, since Q contains no parameters, we need not compute $|Q^+|$, but only make use of $(\tau^2)^{\frac{n-c}{2}}$, but this reasoning justifies setting $|Q^+|/|Q^+| = 1$ in MCMC acceptance ratio calculations.

I'll say a few more words on singular precision and covariance matrices. With a singular precision matrix, there is at least one linear combination, $\Gamma_n^T\theta$, where Γ_n is the last eigenvector, that has zero contribution to the prior density, because $\lambda_n = 0$. The linear combination(s) is ignored. Thus we have a density function but it cannot be integrated because the normalizing constant is infinity, and the prior has nothing to say about the probability density of at least one linear combination. In contrast, with a singular covariance matrix, there is at least one linear combination that has zero variance, thereby imposing a constraint on at least one linear combination of any realization, $\Gamma_n^T\theta \equiv 0$. The constraint on the linear combination must be satisfied, and any θ that does not satisfy this constraint has zero density.

This impropriety carries over into the marginalized model for data. For example, suppose we have $Y \sim \mathcal{N}(X\beta + \theta, V_Y)$, with an MRF prior for θ, $\theta \sim \mathcal{N}(0, \tau^2 Q^{-1})$, again using Q^{-1} only conceptually, and where $X\beta$ contains terms that take the place of the constrained linear combinations. Marginalizing over θ, we have an improper prior predictive distribution for Y, $Y \sim \mathcal{N}(X\beta, \Sigma)$ where $\Sigma = V_Y + \tau^2 Q^{-1}$, but Σ does not exist because the inverse of Q does not exist, with c linear combinations having infinite variance. This improper prior predictive distribution makes sense: we cannot generate from the prior on θ, so we cannot generate from the predictive distribution for the data. However, the posterior for θ will be proper for $n > c$, as will the posterior predictive distribution. Note that in the constrained space, the prior for Y is proper, so we can generate Y with constraints on the appropriate linear combinations by generating realizations of θ with the constraints. This involves using the pseudoinverse to get the predictive distribution, $Y \sim \mathcal{N}_n(X\beta, V_Y + \tau^2 Q^+)$, which is the prior predictive when the linear constraints on θ are imposed in its prior through the pseudoinverse.

Two solutions to identifiability

Solution 1 In most cases we are not concerned with generating from the prior predictive and attention is focused on the posterior for θ and the other parameters. Here it is simplest to work on the precision scale, with zero prior precision (infinite prior variance) on the c linear combinations, and with the linear combinations not included in $X\beta$, including these linear combinations implicitly as part of θ. The conditional posterior for θ has variance $V_{\phi|Y} = (V_Y^{-1} + \tau^{-2}Q)^{-1}$, which is full rank, with the prior contributing no information about the c linear combinations. This is equivalent to putting flat, improper priors on those linear combinations. One can sample from the conditional for θ and the c linear combinations will be unconstrained by the prior, but informed by the data.

In many cases we would integrate over θ. We can express the marginal precision for Y based on completing the square as $\Sigma^{-1} = V_Y^{-1} - V_Y^{-1}(V_Y^{-1} + \tau^{-2}Q)^{-1}V_Y^{-1}$. This precision matrix has c zero eigenvalues, inherited from the improper prior for θ. The normalizing constant for the marginal density of Y can be expressed in terms of $|\Sigma|^{1/2} = |V_Y|^{1/2}|V_{\phi|Y}^{-1}|^{1/2}(\tau^2)^{(n-c)/2}/|Q|^{1/2}$ where $|Q|$ is a constant and need not be computed. Thus the marginalized likelihood can be computed even though the distribution for Y is improper (but proper in a reduced dimension subspace), since we do not need to compute Σ. Note that the marginal likelihood ignores c linear combinations of Y because their precisions are zero, thereby taking these combinations to have no information about the remaining parameters after marginalization.

Solution 2 In some cases, one may wish to enforce the constraint(s) in the prior and then include the constrained linear combinations as parameters in $X\beta$, which is an alternative that also gives identifiability. When sampling θ in an MCMC (potentially off-line if it has been integrated out), one can use one of two approaches to sample θ such that the constraints are obeyed. One solution is to set the last c eigenvalues of $V_{\phi|Y}$ to be zero, while a potentially more efficient approach that allows one to use sparse matrix routines and avoid the eigendecomposition is to use conditioning by kriging as described in Rue and Held (2005, p. 37). A common ad hoc approach is to instead sample θ without constraint and then impose the constraint empirically, for example, setting $\theta^{\text{new}} \equiv \theta - \bar{\theta}1$, for the constraint on the mean. Unfortunately, this is not the same as sampling under the constraint (see Rue and Held (2005, p. 36)), so it does not preserve the posterior as the stationary distribution of the MCMC, although in practice the departure may be minimal.

Enforcing full propriety via an autoregressive parameter

As discussed in Banerjee et al. (2004), one can instead enforce propriety by taking $Q = D^{-1} - \rho W$ with constraints on ρ (the inverses of the largest and smallest eigenvalues of $D^{1/2}WD^{1/2}$). ρ is analogous to the AR(1) autocorrelation parameter, and its presence allows us to generate from the MRF distribution with nonzero $\bar{\theta}$ in the simple CAR setting. However, the interpretation of this model structure is troublesome as the resulting full conditional mean relates to ρ times the average of the neighbors in the simple CAR setting, $\rho \sum_{j \in N(i)} \theta_j / m_i$, causing shrinkage toward the overall mean (zero in my simplified setting here). In a spatial setting, this is generally unappealing, in contrast to the time series context. Also, in many empirical examples the estimate of ρ ends up being nearly one. Hence to my mind, the intrinsic model is more appealing and has a clean interpretation as a proper prior in a subspace, either with enforced constraints or taking an improper prior on the unconstrained linear combinations.

References

Banerjee, S., Carlin, B., and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, Florida: Chapman & Hall.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton: Chapman & Hall.

Yue, Y. and Speckman, P. (2008), Nonstationary spatial Gaussian Markov random fields Technical report, University of Missouri, Department of Statistics.