Small Scale Sim

Morris, S.

July 1, 2015

Small-scale simulation results

Data settings

I generated data at 2000 sites using the following settings:

- $\alpha = 0.3$
- $\rho = 0.1$
- $\xi = 0$

This is 1500 sites for training, and 500 sites for cross validation. The MCMC ran for 45000 iterations with 35000 burnin, and convergence isn't great, but it's more stable than before.

Methods

There were 5 methods used to fit the datasets.

> 1a: Fix α and ρ in the MCMC to be the estimates from the pairwise composite likelihood. Only fit β and random effects in the MCMC.
>
> 1b: Fix *alpha* in the MCMC to be the estimate from the pairwise composite likelihood, and fix *rho* in the MCMC to be the knot spacing. Only fit β and random effects in the MCMC.
>
> 1c: Fit α, ρ, *beta*, and random effect in the MCMC.
>
> 2: Logit
>
> 3: Probit

Methods 1a, 1b, and 1c are all variations on rare binary. Method 1a uses the pairwise composite likelihood to estimate α and ρ. Method 1b uses the pairwise composite likelihood to estimate α with ρ taken as the knot spacing. Both 1a and 1b take β to be the estimate for β marginally for the sites when estimating α and ρ.

Results

Here are the results for 10 datasets:

Table 1: Simulation results (x 100) for 10 datasets

	1a	1b	1c	4	5
1	2.739	2.356	2.331	5.477	2.317
2	2.071	1.881	1.813	1.66	1.752
3	3.932	3.799	3.816	3.745	3.763
4	2.201	1.938	1.929	1.862	2.076
5	1.33	0.9368	0.8578	1.019	0.9547
6	0.6124	0.2163	NA	0.3369	0.3013
7	2.017	1.767	1.757	1.789	1.834
8	4.6	0.7708	0.8166	0.6557	0.6397
9	2.605	2.531	2.564	2.866	2.596
10	1.528	1.147	1.036	1.162	1.153
Mean	2.364	1.734	1.88	2.057	1.739

So, some combination of 1b and 1c tend to help the performance of our rare binary method the most. I also encountered a little bit of problem with dataset 6 when trying to fit all the parameters in the MCMC. When I didn't fix α and ρ, as the MCMC was getting started, eventually the random effects would stop moving in the MCMC. I don't know why they stopped.

Somewhat informative priors

We also wanted to explore what happens when we give an informative prior on α and ρ in the MCMC. For α, in each of the three methods listed below, we used a beta distribution with the mean taken to be the estimate from the pairwise composite likelihood and a standard deviation of 0.05. Each of the methods below differs in how we treat ρ.

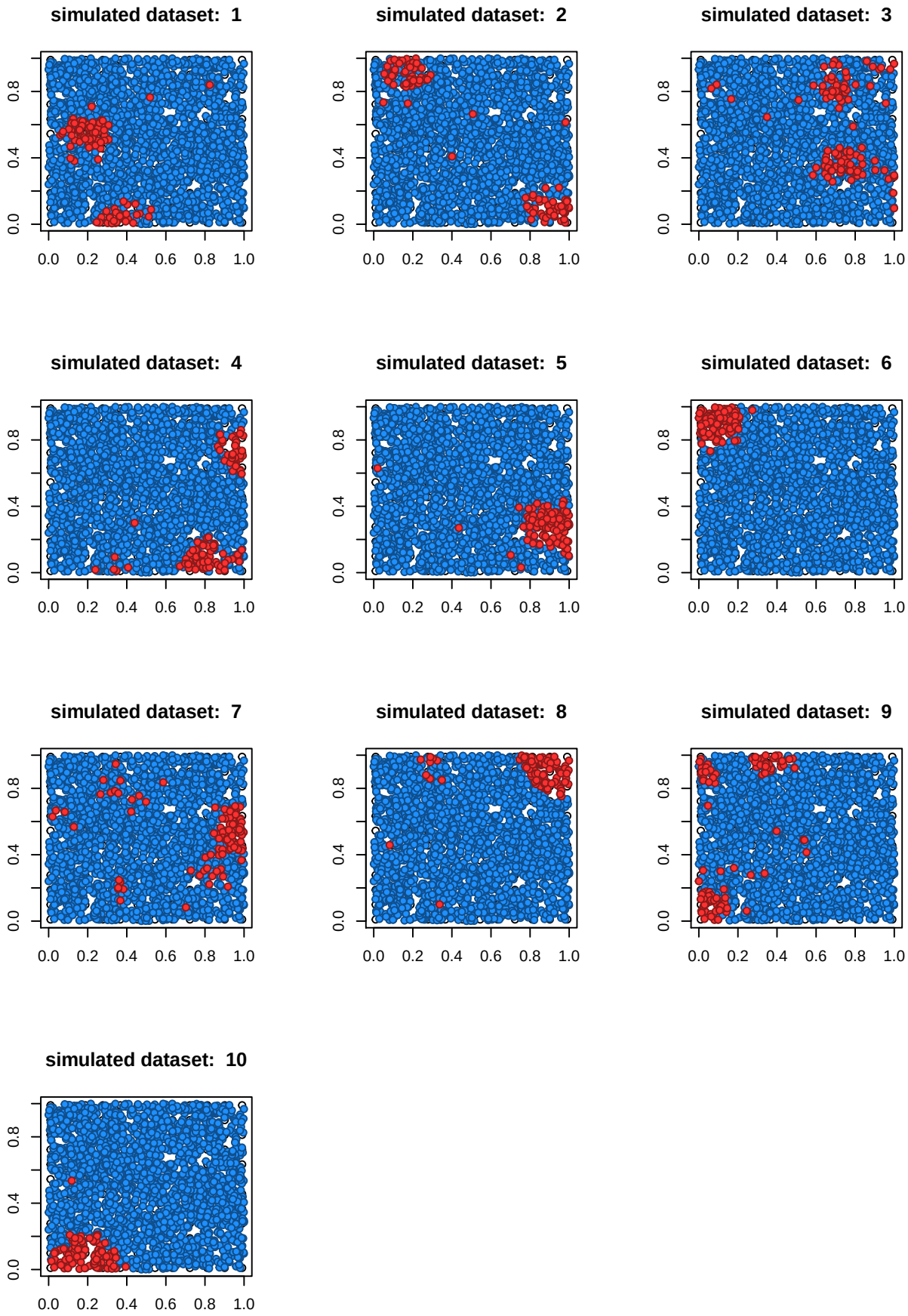
1a: The mean of ρ in the MCMC is set to be the estimates from the pairwise composite likelihood.

1b: The mean of ρ in the MCMC is set to be the knot spacing.

1c: We fix ρ in the MCMC to be the knot spacing.

Table 2: Simulation results (x 100) for 10 datasets

	2a	2b	2c	4	5
1	2.344	2.358	2.367	5.477	2.317
2	1.911	1.857	1.881	1.66	1.752
3	3.822	3.819	3.777	3.745	3.763
4	1.926	1.931	1.943	1.862	2.076
5	0.9076	0.8977	0.8879	1.019	0.9547
6	4.4	0.3167	0.2283	0.3369	0.3013
7	1.75	1.753	1.763	1.789	1.834
8	1.406	1.814	0.9669	0.6557	0.6397
9	2.56	2.566	2.545	2.866	2.596
10	4.4	1.023	1.024	1.162	1.153
Mean	2.543	1.834	1.738	2.057	1.739


simulated dataset: 1
simulated dataset: 2
simulated dataset: 3
simulated dataset: 4
simulated dataset: 5
simulated dataset: 6
simulated dataset: 7
simulated dataset: 8
simulated dataset: 9
simulated dataset: 10
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0

Assessing variability of α due to site inclusion.

I also tried fitting a variety of different max distances for the pairwise likelihood to see how the estimate of α changes as we include sites that are further away. Let h be the knot distance in the x or y direction. I considered including pairs of sites if they were $h, 2h, \ldots, 5h$ away. These are the different estimates of α when ρ is fixed.

Table 3: Pairwise estimates for α

	1 x h	2 x h	3 x h	4 x h	5 x h
1	0.401	0.3754	0.3774	0.3777	0.3777
2	0.4529	0.4114	0.4101	0.4102	0.4102
3	0.7001	0.6483	0.6478	0.6481	0.6481
4	0.327	0.2913	0.2904	0.2906	0.2906
5	0.2105	0.1843	0.1809	0.1809	0.1809
6	0.1125	0.1008	0.1005	0.1005	0.1005
7	0.4509	0.4201	0.4178	0.4179	0.4179
8	0.1122	0.101	0.1005	0.1005	0.1005
9	0.3841	0.369	0.3684	0.3683	0.3683
10	0.178	0.147	0.144	0.1443	0.1443
Mean	0.3329	0.3049	0.3038	0.3039	0.3039

It would appear that the estimates seem to stabilize as long as we include sites in the pairwise composite likelihood that are within $2h$ to $3h$ of the reference site.

Tweaks to likelihood and MCMC

I recoded a few of the functions in C to help with time savings, and I also made two additional changes.

1. Instead of using $\theta^* \frac{1}{z^{1/\alpha}}$ where $\theta^* = \sum_{l=1}^{L} A_l w_l^{1/\alpha}$ when doing predictions at unobserved locations, I'm now using $\sum_{l=1}^{L} A_l \psi_l^*$ where $\psi_l^* = \frac{w_l^{1/\alpha}}{z^{1/\alpha}}$. In the θ^* parameterization, due to the magnitude of the A_l terms, there is more potential for numerical problems when dividing by $z^{1/\alpha}$. The ψ_l^* parameterization is preferred because z and w are closer in magnitude.
2. In preparation for the simulated dataset with more knots, we made two minor tweaks to the likelihood calculcations. We now have a cutoff value for the distance at which A_l no longer impacts z. This impacts the function `updateA` as well as the value for w_l. We automatically set $w_l = 0$ when the location where y is observed is too far away from the knot location.

The methods used are otherwise identical to the methods used in 2a, 2b, and 2c.

Table 4: Simulation results (x 100) for 10 datasets

	2a*	2b*	2c*	4	5
1	2.342	2.363	2.362	5.477	2.317
2	1.919	1.87	1.885	1.66	1.752
3	3.82	3.82	3.787	3.745	3.763
4	1.922	1.946	1.96	1.862	2.076
5	0.8654	0.9128	0.8779	1.019	0.9547
6	0.4021	0.3599	0.2772	0.3369	0.3013
7	1.752	1.762	1.764	1.789	1.834
8	0.6041	0.6294	0.5715	0.6557	0.6397
9	2.548	2.57	2.549	2.866	2.596
10	1.025	1.025	1.055	1.162	1.153
Mean	1.72	1.726	1.709	2.057	1.739

The difference between our method and probit is very small. Although the method that performs the best here, 2c*, fixes ρ at the knot spacing, both methods 2a* and 2b* which do not fix ρ also show some minor improvement over the probit.