STAT647 Homework 2 (Due 3rd Oct)

This homework can be done as a group.

There will be two spatial models we consider in this homework

$$\begin{aligned}\text{Model 1} : Y(s) &= 5+\varepsilon_1(s)+e(s)\\ \text{Model 2} : Y(s) &= 5+\varepsilon_2(s)+e(s)\end{aligned}$$

where $e(s)$ are iid normal random variables with $e(s) \sim N(0,\sigma^2=1)$ and $\text{cov}(\varepsilon_i(s_1),\varepsilon_i(s_2)) = c_i(s_1-s_2;\rho,\gamma)$ with

$$\begin{aligned}c_1(s;\rho,\gamma) &= \gamma^2\exp(-\|s\|_2/\rho)\\ c_2(r;\rho,\gamma) &= \sigma^2\exp\left(-\frac{r}{\rho}\right)\cos\left(r\right), \qquad 0<\rho\leq\tan(\pi/(2d))\end{aligned}$$

where we set $\gamma=2$ and $\rho=0.75$. The first is the exponential covariance the second is the exponential damped covariance (the code in R is for the exponential damped covariance is RMdampedcos, I think). I had made a mistake in my classnotes on how the exponential damped covariance is defined, which I have now corrected. Be a little careful when maximising likelihoods using the exponential damped covariance, in the case that $d=2$, the range parameter is restricted to $0<\rho\leq 1$. If $\rho>1$, the covariance is not positive definite and the function RMdampedcos, may not work (I am not sure).

Run all the simulations with $M=1000$ replications (if your computer allows). If we use too few replications the empirical standard errors and empirical biases that we obtain will not be close to the true population standard errors and empirical biases. The number of replications we use in a simulation study should be independent of the sample sizes used. In all simulations where different estimators are compared use the same set replication (seeds) over all the estimators.

This HW uses the REML, conditional likelihood, pairwise likelihood, regular Whittle likelihood and debiased Whittle likelihood (the last two are only for gridded data) and tapered likelihood approximation. All groups who submitted code, must include a corresponding email address, to whom students can send questions if they have problems running your code.

For all simulations include tables, with empirical bias, standard deviation and histograms of the estimates.

If running the simulations are too difficult on your computer you can access our clusters. Please ask our PhD students for details on this if you are a masters student.

(1) In this question we compare the Guassian maximum likelihood estimator with the REML estimator for the covariance parameters $\gamma = 2$, $\rho = 0.75$ and $\sigma^2 = 1$.

In all the simulations below use both the exponential and the exponential damped covariance.

(a) (i) For $d = 1$, uniformly sample the spatial process over the range $[0, 2]$ and $n = 100$. Estimate the parameters using the GMLE and REML.

(ii) For $d = 1$, uniformly sample the spatial process over the range $[0, 10]$ and $n = 1000$. Estimate the parameters using the GMLE and REML.

(iii) For $d = 1$, uniformly sample the spatial process over the range $[0, 2]$ and $n = 100$. Estimate the parameters using the GMLE and REML.

(iv) For $d = 1$, uniformly sample the spatial process over the range $[0, 10]$ and $n = 1000$. Estimate the parameters using the GMLE and REML.

(b) (i) For $d = 2$, uniformly sample the spatial process over the range $[0, 2]^2$ and $n = 100$. Estimate the parameters using the GMLE and REML.

(ii) For $d = 2$, uniformly sample the spatial process over the range $[0, 2]^2$ and $n = 1000$. Estimate the parameters using the GMLE and REML.

(iii) For $d = 2$, uniformly sample the spatial process over the range $[0, 10]^2$ and $n = 100$. Estimate the parameters using the GMLE and REML.

(iv) For $d = 2$, uniformly sample the spatial process over the range $[0, 10]^2$ and $n = 1000$. Estimate the parameters using the GMLE and REML.

(c) Remember tabulate the simulations in a coherent readable fashion and comment on what you observe.

Are there differences between bias and standard errors if they same locations are used in each replication. Or a different sample of locations (generated using a uniform distribution) is used each time?

(2) We now consider extremely large data sets (where the covariance matrix is computationally expensive to invert; and may even be close to singular). We consider data that is unformly sampled over the region.

In this question, compare the conditional likelihood, pairwise likelihood and tapered likelihood approximation. We use both the exponential and the exponential damped covariance with $\gamma = 2$, $\rho = 0.75$ and $\sigma^2 = 1$.

(a) (i) For $d = 1$, uniformly sample the spatial process over the range $[0, 2]$ and $n = 10000$.

(ii) For $d = 1$, uniformly sample the spatial process over the range $[0, 10]$ and $n = 10000$.

(b) (i) For $d = 2$, uniformly sample the spatial process over the range $[0, 2]^2$ and $n = 10000$.

(ii) For $d = 2$, uniformly sample the spatial process over the range $[0, 10]^2$ and $n = 10000$.

(c) Remember tabulate the simulations in a coherent readable fashion and comment on what you observe.

Are there differences between bias and standard errors if they same locations are used in each replication. Or a different sample of locations (generated using a uniform distribution) is used each time?

If possible report the run (CPU) times.

(3) We now consider extremely large data sets (where the covariance matrix is computationally expensive to invert; and may even be close to singular). We consider data that is sampled from a regular grid.

In this question, compare the conditional likelihood, pairwise likelihood and tapered likelihood approximation. We use both the exponential and the exponential damped covariance with $\gamma = 2$, $\rho = 0.75$ and $\sigma^2 = 1$.

(a) (i) For $d = 1$, sampled on a regular grid $\{2t/n; t = 1, \ldots, n\}$ and $n = 10000$.

(ii) For $d = 1$, sampled on the regular grid $\{10t/n; t = 1, \ldots, n\}$ and $n = 10000$.

(b) (i) For $d = 2$, sampled on a regular grid $\{(2i/m, 2j/m); 1 \leq i, j \leq m\}$ and $m^2 = 10000$.

(ii) For $d = 2$, sampled on a regular grid $\{(10i/m, 10j/m); 1 \leq i, j \leq m\}$ and $m^2 = 10000$.

(c) Remember tabulate the simulations in a coherent readable fashion and comment on what you observe.

Are there differences between bias and standard errors if they same locations are used in each replication. Or a different sample of locations (generated using a uniform distribution) is used each time?

If possible report the run (CPU) times.