Lecture 5: Geostatistics

Dennis Sun

July 17, 2014

1 Review from Last Lecture

1.1 An Outline So Far

We started off this class by considering the classical statistical model

$$y_i = \mu_i + \epsilon_i$$

where μ_i is typically some function of some covariates (i.e., $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$) and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. We decided that this was unsatisfactory for spatial and temporal data because there may be correlation between the errors across times or locations.

We allowed for spatial dependence by introducing an autoregressive term:

$$y_i - \mu_i = \phi \sum_{j=1}^{n} w_{ij}(y_j - \mu_j) + \epsilon_i.$$

Estimation is still simple: we alternate between estimating $\boldsymbol{\mu}$ and estimating ϕ. For time series, least-squares can be used to estimate ϕ, but in general, we need to use maximum likelihood.

However, there is a weakness of this model: it assumes that the y_i are observed without noise. We can extend this to a more general state-space model:

$$\begin{aligned} &\textit{State model}: && y_t - \mu_t = \phi \sum_{j=1}^{n} w_{ij}(y_{t-1} - \mu_{t-1}) + \epsilon_t \\ &\textit{Data model}: && z_t = a y_t + \delta_t \end{aligned}$$

1.2 Kalman Filter

The Kalman filter provided a way to do state estimation without inverting large matrices.

$$\begin{aligned} &\textit{State model}: && y_t - \mu_t = \phi(y_{t-1} - \mu_{t-1}) + \epsilon_t \\ &\textit{Data model}: && z_t = a y_t + \delta_t \end{aligned}$$

Let $y_{t|s}$ denote the best predictor of y_t given $z_1, ..., z_s$, i.e.,

$$\begin{aligned} y_{t|s} &\overset{def}{=} \mathrm{E}(y_t | z_1, ..., z_s) \\ V_{t|s} &\overset{def}{=} \mathrm{Var}(y_t | z_1, ..., z_s) \end{aligned}$$

The Kalman filter iterates between

1. calculating $(y_{t|t-1}, V_{t|t-1})$ from $(y_{t-1|t-1}, V_{t-1|t-1})$, and
2. calculating $(y_{t|t}, V_{t|t})$ from $(y_{t|t-1}, V_{t|t-1})$.

The first of these can be accomplished easily:

$$\begin{aligned} y_{t|t-1} = \mathrm{E}(y_t|z_1,...,z_{t-1}) &= \mu_t + \phi(\mathrm{E}(y_{t-1}|z_1,...,z_{t-1}) - \mu_{t-1}) \\ &= \mu_t + \phi(y_{t-1|t-1} - \mu_{t-1}) \\ V_{t|t-1} = \mathrm{Var}(y_t|z_1,...,z_{t-1}) &= \phi^2 \mathrm{Var}(y_{t-1}|z_1,...,z_{t-1}) + \sigma^2 \\ &= \phi^2 V_{t-1|t-1} + \sigma^2 \end{aligned}$$

The second task is a bit trickier. However, using the joint distribution

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} \mid z_1,...,z_{t-1} \sim N\left(\begin{pmatrix} y_{t|t-1} \\ a y_{t|t-1} \end{pmatrix}, \begin{pmatrix} V_{t|t-1} & aV_{t|t-1} \\ aV_{t|t-1} & a^2 V_{t|t-1} + \tau^2 \end{pmatrix} \right)$$

we can use the conditional distribution formulas to calculate

$$\begin{aligned} y_{t|t} = \mathrm{E}(y_t|z_1,...,z_t) &= y_{t|t-1} + aV_{t|t-1}(a^2 V_{t|t-1} + \tau^2)^{-1}(z_t - a y_{t|t-1}) \\ V_{t|t} = \mathrm{Var}(y_t|z_1,...,z_t) &= V_{t|t-1} - a^2 V_{t|t-1}^2 (a^2 V_{t|t-1} + \tau^2)^{-1} \end{aligned}$$

1.3 Kalman Smoother

The Kalman filter gives us a way to estimate the state y_t using only the observations up to time t, i.e., $y_{t|t} = \mathrm{E}(y_t|z_1,...,z_t)$. This is useful in real-time implementations, where we have to make estimates of $y_1, y_2, ...$ in real time as data $z_1, z_2, ...$ stream in.

However, if we already have all the data in front of us, we might want an estimator that makes use of all of the data $z_1,...,z_n$. The MMSE estimator in this case is $y_{t|n} = \mathrm{E}(y_t|z_1,...,z_n)$. Is there a similar algorithm that lets us calculate $y_{t|n}$ recursively?

It turns out that there is! It is called the Kalman smoother. You will derive this on your homework. The basic idea is to first derive $\mathrm{E}(y_t|y_{t+1}, z_1,...,z_t)$ and $\mathrm{Var}(y_t|y_{t+1}, z_1,...,z_t)$ using the distribution of

$$\begin{pmatrix} y_t \\ y_{t+1} \end{pmatrix} \mid z_1,...,z_t,$$

and then to calculate

$$\begin{aligned} \mathrm{E}(y_t|z_1,...,z_n) &= \mathrm{E}(\mathrm{E}(y_t|y_{t+1}, z_1,...,z_n)|z_1,...,z_n) \\ &= \mathrm{E}(\mathrm{E}(y_t|y_{t+1}, z_1,...,z_t)|z_1,...,z_n) \end{aligned}$$

and $\mathrm{Var}(y_t|z_1,...,z_n)$ in a similar way.

2 Geostatistics / Kriging

The goal of geostatistics, or kriging, is to predict the value $y_0 \equiv y(s_0)$ of some process (e.g., mineral) at a location s^*, given the values at locations $y_i \equiv y(s_i)$, $i = 1,...,n$.

Typically, we will assume that $y(s)$ is a Gaussian process, which is completely specified by a mean function $\mu(s)$ and covariance function $\Sigma(s,t) \stackrel{def}{=} \mathrm{Cov}(y(s), y(t))$.

As before, we will often assume $\mu(s) = \boldsymbol{x}(s)^T\boldsymbol{\beta}$ (or some other functional form), and we will assume that we have already de-meaned the process, i.e., $y(s) - \mu(s)$, so in what follows, we will assume that $y(s)$ is a zero-mean process, i.e., $\mathrm{E}(y(s)) = 0$ for all s.

2.1 No measurement error

For predicting y_0, it makes sense to find an estimator $f(\boldsymbol{y})$ that minimizes the mean squared error (MSE), i.e.,

$$\mathrm{MSE}_f(y_0) = \mathrm{E}[y_0 - f(\boldsymbol{y})]^2.$$

Theorem 1. *The choice of f that minimizes $MSE_f(y_0)$ is*

$$f(\boldsymbol{y}) = E(y_0|\boldsymbol{y}).$$

Proof.

$$\begin{aligned}\mathrm{E}[y_0 - f(\boldsymbol{y})]^2 &= \mathrm{E}[y_0 - \mathrm{E}(y_0|\boldsymbol{y}) + \mathrm{E}(y_0|\boldsymbol{y}) - f(\boldsymbol{y})]^2 \\ &= \mathrm{E}[y_0 - \mathrm{E}(y_0|\boldsymbol{y})]^2 + \mathrm{E}[\mathrm{E}(y_0|\boldsymbol{y}) - f(\boldsymbol{y})]^2 + 2\underbrace{\mathrm{E}[(y_0 - \mathrm{E}(y_0|\boldsymbol{y}))(\mathrm{E}(y_0|\boldsymbol{y}) - f(\boldsymbol{y}))]}_{0}.\end{aligned}$$

In this form, we can clearly see that the MSE is minimized by setting $f(\boldsymbol{y}) = \mathrm{E}(y_0|\boldsymbol{y})$ so that the second term is zero. □

Now, since $\begin{pmatrix}\boldsymbol{y} \\ y_0\end{pmatrix}$ is multivariate Gaussian, we can calculate the MMSE estimator $\mathrm{E}(y_0|\boldsymbol{y})$ using the usual multivariate normal formulas.

Theorem 2. *The MMSE predictor for y_0 is*

$$\hat{y}_0 = E(y_0|\boldsymbol{y}) = \Sigma_{y_0,\boldsymbol{y}}\Sigma_{\boldsymbol{yy}}^{-1}\boldsymbol{y},$$

where $\Sigma_{\boldsymbol{yy}}$ denotes the $n \times n$ matrix obtained by evaluating the covariances at each of the y_i, and $\Sigma_{y_0,\boldsymbol{y}} \overset{def}{=} \begin{bmatrix}E(y_0y_1) & \cdots & E(y_0y_n)\end{bmatrix}$ denotes the covariance of y_0 with each of the y_i.

Proof. This follows from

$$\begin{pmatrix}\boldsymbol{y} \\ y_0\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{0} \\ 0\end{pmatrix}, \begin{pmatrix}\Sigma_{\boldsymbol{yy}} & \Sigma_{\boldsymbol{y},y_0} \\ \Sigma_{y_0,\boldsymbol{y}} & \Sigma_{y_0,y_0}\end{pmatrix}\right)$$

and using the conditional expectation formula for multivariate Gaussians. □

2.2 Alternative Derivation without Assuming Normality

In many applications, we are not willing to believe the normal assumption. However, the estimator above is still the best linear unbiased predictor, i.e., it is the best estimator of the form $f(\boldsymbol{y}) = \sum_{i=1}^n \alpha_i y_i = \boldsymbol{\alpha}^T\boldsymbol{y}$ that minimizes

$$\begin{aligned}\mathrm{E}\left(y_0 - \sum_{i=1}^n \alpha_i y_i\right)^2 &= \mathrm{E}(y_0^2) - 2\sum_{i=1}^n \alpha_i \mathrm{E}(y_0y_i) + \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\mathrm{E}(y_iy_j) \\ &= \Sigma_{y_0y_0} - 2\Sigma_{y_0,\boldsymbol{y}}\alpha + \boldsymbol{\alpha}^T\Sigma_{\boldsymbol{yy}}\boldsymbol{\alpha}\end{aligned}$$

By differentiating with respect to $\boldsymbol{\alpha}$, setting equal to zero, and solving, we see that the optimal weights are

$$\hat{\boldsymbol{\alpha}}^T = \Sigma_{y_0,\boldsymbol{y}}\Sigma_{\boldsymbol{yy}}^{-1}.$$

Note that when we calculate $\hat{y}_0 = \hat{\boldsymbol{\alpha}}^T\boldsymbol{y}$, we get exactly the same estimator as we did in the previous subsection. This is the approach that is taken by Chapter 2 of the Sherman reference.

2.3 With measurement error

Now suppose we don't observe y_i directly, but only $z_i = y_i + \delta_i$, where $\delta_i \sim N(0, \tau^2)$. Then the MMSE estimator is $\mathrm{E}(y_0|\boldsymbol{z})$. Since

$$\begin{pmatrix}\boldsymbol{z}\\ y_0\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}, \begin{pmatrix}\Sigma_{\boldsymbol{yy}} + \tau^2 I & \Sigma_{\boldsymbol{y},y_0}\\ \Sigma_{y_0,\boldsymbol{y}} & \Sigma_{y_0y_0}\end{pmatrix}\right),$$

the new MMSE (or BLUP, if we're not assuming normality) estimator is

$$\hat{y}_0 = \mathrm{E}(y_0|\boldsymbol{y}) = \Sigma_{y_0,\boldsymbol{y}}(\Sigma_{\boldsymbol{yy}} + \tau^2 I)^{-1}\boldsymbol{y},$$