Summary of Recent Progress Applying MMF to Gaussian Process Regression

Jon Eskreis-Winkler

University of Chicago

eskreiswinkler@uchicago.edu

November 20, 2015

Overview

1. What is a Gaussian Process?
   - Computational Bottleneck
2. Potential for MMF to Help
3. Application: Irish Wind Data
4. Conclusion

Gaussian Process motivation and definition

- Given data $(\mathbf{X}_n, \mathbf{Y}_n) = (X_1, Y_1), \ldots, (X_n, Y_n)$, how do you infer what type of process f generated the data and responses? How do you enable predictions for future data?
- The Bayesian approach uses the posterior probability distribution for the process $f : \mathcal{X}^n \to \mathcal{Y}^n$ is

$$P(f|\mathbf{X}_n, \mathbf{Y}_n) = \frac{P(\mathbf{Y}_n|f, \mathbf{X}_n)P(f)}{P(\mathbf{Y}_n|\mathbf{X}_n)}$$

- What is the function space to which f belongs? Need to restrict the space if we want to be able to measure the numerator terms.
- Simplifying assumption: f is a Gaussian Process.
- A mapping f is called a **Gaussian Process** if for any $m \in \mathbb{N}$, for any m-length input set $(X_1, \ldots, X_m)$, the distribution of $(f(X_1), \ldots, f(X_m))$ is multivariate normal with $(f(X_1), f(X_2), \ldots, f(X_n)) \sim N(\mu(x), K(x))$ with Mercer kernel K.

Gaussian Process Intuition

- Gaussian Process generalizes the notion of a finite dimensional Gaussian distribution to an infinite dimensional analog

$$X \sim \mathcal{N}(\mu, \Sigma) \iff f(\mathbf{X}_n) \sim \mathcal{N}(\mu(\mathbf{X}_n), \Sigma(\mathbf{X}_n))$$

  The mean and covariance functions are defined by the data, assume $\mu(x) = 0$ for simplicity.

- For prediction, the covariance function $\Sigma_p(\mathbf{X}_n)$ is often of the form $\Sigma(\mathbf{X}_n) = K + \sigma^2 I$.

- The prior for f is the density for an n-dimensional multivariate normal:

$$P(f(\mathbf{X}_n)) = (2\pi)^{-\frac{n}{2}} \det(\Sigma(\mathbf{X}_n))^{-\frac{1}{2}} \exp\left(-\frac{1}{2} f(\mathbf{X}_n)^T \Sigma(\mathbf{X}_n)^{-1} f(\mathbf{X}_n)\right)$$

- The likelihood of the data is the conditional distribution $\mathbf{Y}_n|\mathbf{X}_n \sim N(0, K(x) + \sigma^2 I)$ and log-likelihood is

$$\log P(\mathbf{Y}_n|\mathbf{X}_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det|K + \sigma^2 I| - \frac{1}{2}\mathbf{Y}_n^T (K + \sigma^2 I)^{-1}\mathbf{Y}_n$$

- To make predictions, we define the distribution for $Y_{n+1}|\mathbf{X}_n, \mathbf{Y}_n, X_{n+1}$

$$Y_{n+1}|\mathbf{X}_n, \mathbf{Y}_n, X_{n+1} \sim N(\tilde{\mu}(x), \tilde{\Sigma}(x))$$

where $\tilde{\mu} = k(X_{n+1}, \mathbf{X}_n)(K(x) + \sigma^2 I)^{-1}\mathbf{Y}_n$ and
$\tilde{\Sigma} = k(X_{n+1}, X_{n+1}) - k(X_{n+1}, \mathbf{X}_n)(K(x) + \sigma^2 I)^{-1} k(\mathbf{X}_n, X_{n+1})$.

Computational Bottleneck

Common computational complexities

- Using Gauss-Jordan elimination, matrix inversion has complexity $O(n^3)$.
- Finding determinant of a matrix by LU decomposition has complexity $O(n^3)$.

- For large n, this makes evaluation of the log-likelihood of the posterior and predictions using the posterior intractable – recall the inverse and determinant in the log-likelihood function and the inverse involved in prediction.
- What can be done?

MMF Background

- The MMF of a symmetric matrix K is a multi-level factorization that approximates A
$$A \approx Q_1^T \cdots Q_L^T H Q_L \cdots Q_1$$
- H is a matrix which is approximately diagonal – it is diagonal except for a "small" block of nonzero entries.
- $Q_1, Q_2, \ldots, Q_L$ are increasingly local rotations matrices, with a shrinking "active set" of rotations.

Application of MMF to GPR Bottleneck

Suppose that a MMF has been constructed for the matrix $(K + \sigma^2 I)$

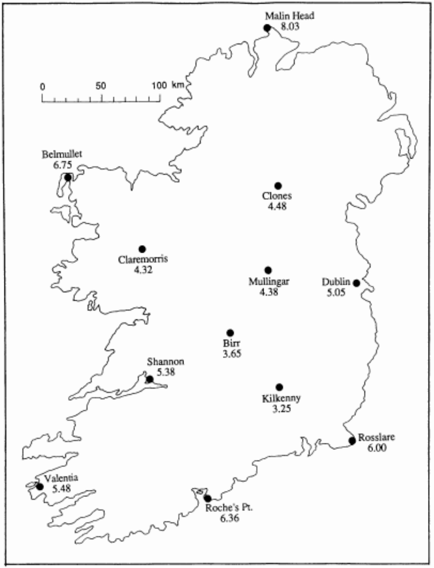
- Determinant computation is only as complicated as computing the determinant of H. The computational cost is $O(nh^2)$ where h is the dimension of H's non-diagonal block. Also, $\forall i \in [1, L],\ \det(Q_i) = 1$.
- Matrix inversion is similarly simplified to $O(nh^2)$ because $\forall i \in [1, L],\ Q_i Q_i^T = I$ and H^{-1} is reduced to inverting the non-diagonal block of H and setting the remaining diagonal elements d_{ii} of H to be $\frac{1}{d_{ii}}$.
- Similar to MMF, these procedures will thus scale roughly linearly with n, assuming h is sufficiently small.

Irish Wind Data

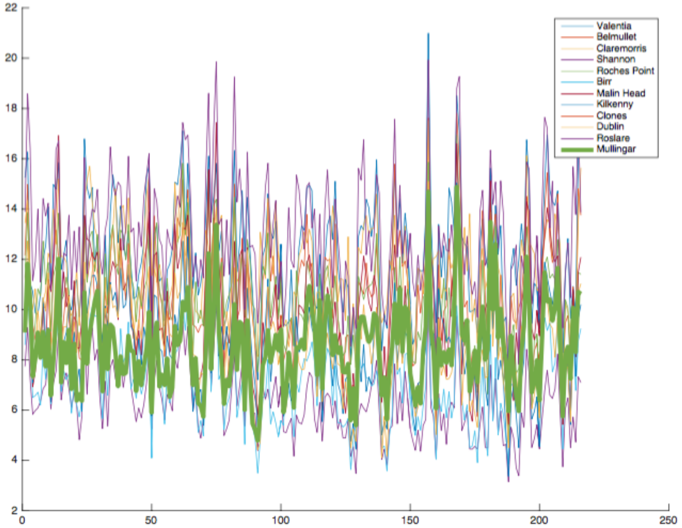
- Daily average wind speeds for January 1961- December 1978 at 12 synoptic meteorological stations in the Republic of Ireland.
- Excellent database for studying spacio-temporal aspects of a dataset.
- Good testbed for potential of MMF. There are (12 sites)(17 years)(365 days) $=$ 74,460 total observations.

Haslett, J. and Raftery, A. E. (1989). Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource (with Discussion). Applied Statistics 38, 1-50.

Irish Wind Data



Irish Wind Data



Can we infer one station's measurements from the others'?

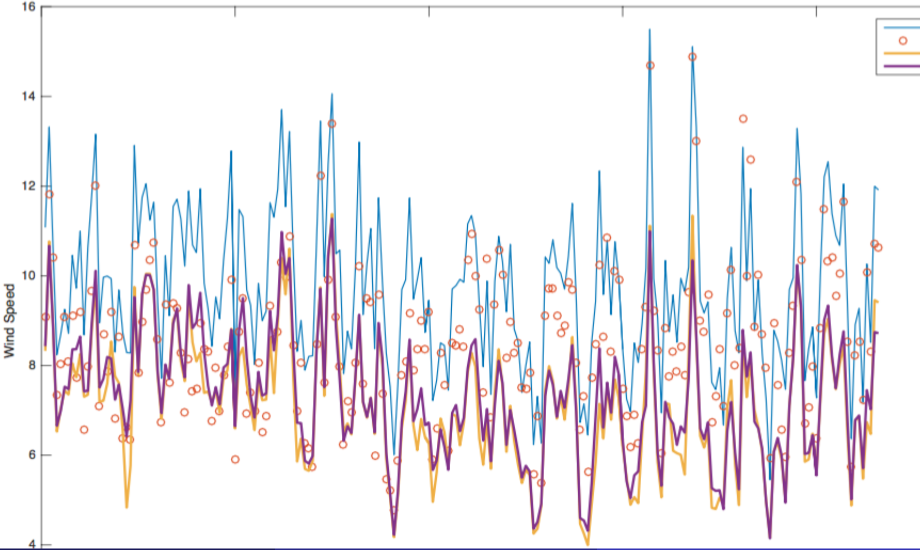
- Training data are 17 years of measurements from all non-Mulligar stations
- Test data is the Mullingar station's measurements.
- To permit comparison with Matlab, needed to shrink data – averaged over months.
- 6 predictor variables: MonthID, Year, Month, Longitude, Lattitude, isCoastal (binary). Response is average monthly wind speed.
- Use RBF kernel for Kernel matrix with varying length-scales:

$$K_{ij} = k(x_i, x_j) + \sigma_n^2 \delta_{i,j} = \sigma_f^2 \exp(-\sum_{k=1}^{6} \frac{(x_{i,k} - x_{j,k})^2}{l_i^2}) + \sigma_n^2 \delta_{i,j}$$

Optimization

- Need to find $\theta = (\sigma_f^2, \sigma_n^2, l)$ to maximize the log-likelihood of $p(\mathbf{Y}_n | \mathbf{X}_n)$.
- Gradient descent is used to maximize log-likelihood, but used direct inverse because MMF determinant option was not yet available.
- Performed a parameter search for θ_{opt} by starting gradient descent on log-likelihood 20 times uniformly over a predefined "reasonable" parameter space. Procedure produced $\{\hat{\theta}^{(i)}\}_{i=1}^{20}$.
- The predictions for all $\theta \in \{\hat{\theta}^{(i)}\}_{i=1}^{20}$ seemed very close so I simply averaged over their predictions.

GPR Performance of MMF



Next Step

- Improve parameter estimates by adding random restarts and by sampling more finely over the parameter space using a cluster.
- Incorporate MMF into the parameter search – not just prediction.
- Develop connection between compression and prediction accuracy.