# Scalable Gaussian Processes with Markovian Covariances

Xiaowei Zhang

Stanford MS&E, December 1, 2017

Hong Kong University of Science and Technology

# Table of Contents

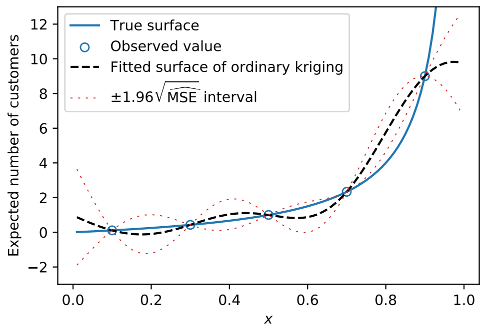
# Gaussian Process (GP)

# Gaussian Process

- Generalization of multivariate normal distribution
  - $(\mathsf{M}(\boldsymbol{x}_1), \ldots, \mathsf{M}(\boldsymbol{x}_n))$ is multivariate normal
- Completely characterized by covariance function $k(\boldsymbol{x}, \boldsymbol{y})$
- Unknown function surface is viewed as a GP realization

$$\mathsf{Z}(\boldsymbol{x}) = \beta + \mathsf{M}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$$

- Leverage *spatial correlation* for prediction



# Applications

- Spatial statistics
  - kriging
- Design and analysis of computer experiments
  - efficient global optimization
- Stochastic simulation
  - stochastic kriging
  - simulation optimization
- Machine learning
  - Gaussian process regression
  - Bayesian optimization

# The Big $n$ Problem

- Function surface is observed at $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with noise

$$z_i = \beta + M(\boldsymbol{x}_i) + \varepsilon(\boldsymbol{x}_i)$$

- The best linear unbiased predictor (BLUP) of $Z(\boldsymbol{x}_0)$ is

$$\widehat{Z}(\boldsymbol{x}_0) = \beta + \boldsymbol{\Sigma}_M(\boldsymbol{x}_0, \cdot)[\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon]^{-1}[\bar{\boldsymbol{z}} - \beta \mathbf{1}_n]$$

- Slow: $[\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon] \in \mathbb{R}^{n \times n}$ and inverting it takes $\mathcal{O}(n^3)$ time
- Numerically instable: $[\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon]$ is often nearly singular

# Approximation Schemes

- Well developed in spatial statistics and machine learning
  - Banerjee, Carlin, and Gelfand (2015)
  - Rasmussen and Williams (2006)
- Reduced-rank approximations
  - Emphasize large-scale dependences, but fail to capture small-scale dependences
- Sparse approximations (e.g., covariance tapering)
  - Emphasize small-scale dependences, but fail to capture large-scale dependences

# Markovian Structure and Sparsity

# Gaussian Markov Random Field (GMRF)

- Specify $\mathbf{\Sigma}_{\mathrm{M}}^{-1}$, instead of $k(\cdot, \cdot)$
- Use graph to describe Markovian structure
  - Given all its neighbors, node $i$ is *conditionally independent* of its non-neighbors
  - E.g., $\mathrm{M}(\boldsymbol{x}_2) \perp (\mathrm{M}(\boldsymbol{x}_0), \mathrm{M}(\boldsymbol{x}_4))$, given $(\mathrm{M}(\boldsymbol{x}_1), \mathrm{M}(\boldsymbol{x}_3))$


- Sparsity: $\mathbf{\Sigma}_{\mathrm{M}}^{-1}(i, j) \neq 0 \iff i$ and $j$ are neighbors
- The sparsity can reduce necessary computation to $\mathcal{O}(n^2)$

# Disadvantages of GMRF

- Hard to specify desired correlation behavior
- Cannot predict locations "off the grid"

$$\widehat{Z}(x_0) = \beta + \underbrace{\boldsymbol{\Sigma}_M(x_0, \cdot)}_{\text{unknown}}[\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon]^{-1}[\bar{z} - \beta \mathbf{1}_n]$$

- Continuous design space must be discretized first, which may result in $N \gg n$ grid points
  - Computing predictor requires $\mathcal{O}(N^2)$
  - $\mathcal{O}(N^2)$ v.s. $\mathcal{O}(n^3)$?

# The Best of Two Worlds?

- We will construct a class of covariance functions for which:
    1. $\mathbf{\Sigma}_M$ can be inverted analytically
    2. $\mathbf{\Sigma}_M^{-1}$ is sparse
- Explicit link between covariance function and sparsity

# Complexity Reduction

- Woodbury matrix identity

$$[\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon]^{-1} = \underbrace{\boldsymbol{\Sigma}_M^{-1}}_{\text{known}} + \underbrace{\boldsymbol{\Sigma}_M^{-1}}_{\text{sparse}} \left[ \underbrace{\boldsymbol{\Sigma}_M^{-1} + \boldsymbol{\Sigma}_\varepsilon^{-1}}_{\text{sparse}} \right]^{-1} \boldsymbol{\Sigma}_M^{-1}$$

  - inversion: $\mathcal{O}(n^2)$
  - multiplications: $\mathcal{O}(n^2)$
  - addition: $\mathcal{O}(n^2)$
- It takes $\mathcal{O}(n^2)$ time to compute BLUP

$$\widehat{Z}(\boldsymbol{x}_0) = \beta + \underbrace{\boldsymbol{\Sigma}_M(\boldsymbol{x}_0, \cdot)}_{\text{known}} [\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\varepsilon]^{-1} [\bar{\boldsymbol{z}} - \beta \mathbf{1}_n]$$

  - If the noise is negligible ($\boldsymbol{\Sigma}_\varepsilon \approx \mathbf{0}$), then no numerical inversion is needed and computing BLUP is $\mathcal{O}(n)$!

# Stability Improvement

1. $\mathbf{\Sigma}_M$ can be made much better conditioned
2. Woodbury also improves numerical stability

$$[\mathbf{\Sigma}_M + \mathbf{\Sigma}_\varepsilon]^{-1} = \mathbf{\Sigma}_M^{-1} + \mathbf{\Sigma}_M^{-1}\left[\mathbf{\Sigma}_M^{-1} + \mathbf{\Sigma}_\varepsilon^{-1}\right]^{-1}\mathbf{\Sigma}_M^{-1}$$

- The diagonal entries of $\mathbf{\Sigma}_\varepsilon^{-1}$ are often large

# 1-D Markovian Gaussian Processes

- Assume $\mathcal{X} = [0, 1]$ and $x_i = \frac{i}{n+1}$, $i = 1, \ldots, n$
- Brownian motion: $k(x, y) = \min(x, y)$

$$\mathbf{\Sigma}_M^{-1} = (n+1) \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}$$

- Brownian bridge: $k(x, y) = \min(x, y) - xy$

$$\boldsymbol{\Sigma}_M^{-1} = (n+1) \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

- Ornstein-Uhlenbeck process: $k(x,y) = \frac{\sigma^2}{2\theta} e^{-\theta|x-y|}$

$$\mathrm{d}X(t) = -\theta X(t)\,\mathrm{d}t + \sigma \mathrm{d}B(t)$$

$$\boldsymbol{\Sigma}_M^{-1} = \frac{\theta}{\sigma^2 \sinh(\theta h)} \begin{pmatrix} e^{\theta h} & -1 & & & \\ -1 & 2\cosh(\theta h) & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2\cosh(\theta h) & -1 \\ & & & -1 & e^{\theta h} \end{pmatrix},$$

with $h = 1/(n+1)$

# Key Observation

- Share the same functional form

$$k(x,y) = p(x)q(y)\,\mathbb{I}_{\{x\le y\}} + p(y)q(x)\,\mathbb{I}_{\{x>y\}},$$

  for some functions $f$ and $g$
- $k_{\mathrm{BM}}(x,y) = x\,\mathbb{I}_{\{x\le y\}} + y\,\mathbb{I}_{\{x>y\}}$
- $k_{\mathrm{BR}}(x,y) = x(1-y)\,\mathbb{I}_{\{x\le y\}} + y(1-x)\,\mathbb{I}_{\{x>y\}}$
- $k_{\mathrm{OU}}(x,y) = e^{x}e^{-y}\,\mathbb{I}_{\{x\le y\}} + e^{y}e^{-x}\,\mathbb{I}_{\{x>y\}}$

# Key Observation

- Share the same functional form

$$k(x,y) = p(x)q(y)\,\mathbb{I}_{\{x\le y\}} + p(y)q(x)\,\mathbb{I}_{\{x>y\}},$$

for some functions $f$ and $g$

- $k_{\mathrm{BM}}(x,y) = x\,\mathbb{I}_{\{x\le y\}} + y\,\mathbb{I}_{\{x>y\}}$
- $k_{\mathrm{BR}}(x,y) = x(1-y)\,\mathbb{I}_{\{x\le y\}} + y(1-x)\,\mathbb{I}_{\{x>y\}}$
- $k_{\mathrm{OU}}(x,y) = e^x e^{-y}\,\mathbb{I}_{\{x\le y\}} + e^y e^{-x}\,\mathbb{I}_{\{x>y\}}$

**Theorem (Ding and Z, 2017)**

If $K$ is nonsingular, then $K^{-1}$ is tridiagonal.

- $\{x_1, \ldots, x_n\}$ are not necessarily equally spaced

# Proof by Linear Algebra

- Show $(K^{-1})_{i,j} = 1$ for $|j - i| \geq 2$ by induction on $n$
- Relation between the inverse and the minors

$$(K)^{-1}_{i,j} = \frac{1}{|K|}(-1)^{i+j} M_{j,i}$$

- Use the Laplace expansion of the determinant

$$|K| = \sum_{\ell=1}^{n} (-1)^{i+\ell} k_{i,l} M_{i,l} = \sum_{\ell=1}^{n} (-1)^{\ell+j} k_{\ell,j} M_{\ell,j}$$

# Positive Definiteness

- What conditions make $K$ positive definite (PD)?

## Positive Definiteness

- What conditions make $K$ positive definite (PD)?

**Theorem (Ding and Z, 2017)**

1. For $n \geq 2$, $|K| = p(x_1)q(x_n)\prod_{i=1}^{n}[p(x_i)q(x_{i-1}) - p(x_{i-1})q(x_i)]$.

# Positive Definiteness

- What conditions make $\boldsymbol{K}$ positive definite (PD)?

**Theorem (Ding and Z, 2017)**

1. For $n \geq 2$, $|\boldsymbol{K}| = p(x_1)q(x_n)\prod_{i=1}^{n}[p(x_i)q(x_{i-1}) - p(x_{i-1})q(x_i)]$.
2. $k(x, y)$ is PD *if and only if*
    - $p(x)q(y) - p(y)q(x) < 0$ for all $x < y$;
    - $p(x)q(y) > 0$ for all $x, y$.

- We call such $k(x, y)$ (1-dimensional) Markovian covariance function (MCF)

# A Naïve Example

- Let $q(x) \equiv 1$ and $p(x)$ be positive, strictly increasing
  - $k(x, y) = p(x)\,\mathbb{I}_{\{x \le y\}} + p(y)\,\mathbb{I}_{\{x > y\}}$, not very reasonable
  - $k(x, y) = \min(p(x), p(y))$: “time-changed” Brownian motion

# A Naïve Example

- Let $q(x) \equiv 1$ and $p(x)$ be positive, strictly increasing
    - $k(x, y) = p(x)\,\mathbb{I}_{\{x \leq y\}} + p(y)\,\mathbb{I}_{\{x > y\}}$, not very reasonable
    - $k(x, y) = \min(p(x), p(y))$: "time-changed" Brownian motion

- How to construct MCFs easily?

The Green’s function of a Sturm-Liouville equation has exactly the same form!

# Sturm-Liouville (S-L) Problem

# S-L Differential Equation

- Assume $\mathcal{X} = [L, R]$

$$\mathcal{L}f(x) \coloneqq \frac{1}{w(x)}\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-u(x)\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right) + v(x)f(x)\right] = h(x),$$

with boundary condition

$$\begin{cases} \alpha_L f(L) + \beta_L f'(L) = 0, \\ \alpha_R f(R) + \beta_R f'(R) = 0. \end{cases}$$

- Any second-order linear ordinary differential equations can be recast in the form of the S-L equation

# Green's Function

- The Green's function satisfies $\mathcal{L}G(x, y) = \delta(x - y)$ with the same BC
- The S-L solution is $f(x) = \int_L^R G(x, y)h(y)\mathrm{d}y$
- Form of Green's function:

$$G(x, y) = cf_1(x)f_2(y)\,\mathbb{I}_{\{x \le y\}} + cf_1(y)f_2(x)\,\mathbb{I}_{\{x > y\}},$$

for some constant $c$, where $\mathcal{L}f_i = 0$, with

$$\alpha_L f_1(L) + \beta_L f_1'(L) = 0 \quad \text{and} \quad \alpha_R f_2(R) + \beta_R f_2'(R) = 0.$$

# Green's Function

- The Green's function satisfies $\mathcal{L}G(x,y) = \delta(x-y)$ with the same BC
- The S-L solution is $f(x) = \int_L^R G(x,y)h(y)\mathrm{d}y$
- Form of Green's function:

$$G(x,y) = cf_1(x)f_2(y)\,\mathbb{I}_{\{x \le y\}} + cf_1(y)f_2(x)\,\mathbb{I}_{\{x > y\}},$$

  for some constant $c$, where $\mathcal{L}f_i = 0$, with

$$\alpha_L f_1(L) + \beta_L f_1'(L) = 0 \quad \text{and} \quad \alpha_R f_2(R) + \beta_R f_2'(R) = 0.$$

- What kind of S-L equations yield a PD Green's function?

# The Eigenvalue Problem (S-L Problem)

- $\mathcal{L}f(x) = \lambda f(x)$ with the same boundary condition
- *Regular*: $u, u', v, w$ are continuous, and $u, w > 0$

**S-L Theory**

The regular S-L problem has a countable number of real eigenvalues and the normalized eigenfunctions can be chosen real-valued and form an orthonormal basis.

- If the eigenvalues are all positive, then
  - $G(x, y) = \sum_\ell \lambda_\ell^{-1} \phi_\ell(x) \phi_\ell(y)$
  - PSD

# Positive Eigenvalues

- Integration by parts

$$\lambda_\ell = \langle \mathcal{L}\phi_\ell, \phi_\ell \rangle = - u(x)\phi_\ell(x)\phi_\ell'(x)\Big|_{x=L}^{R} + \int_L^R u(x)[\phi_\ell'(x)]^2 \mathrm{d}x + \int_L^R v(x)[\phi_\ell(x)]^2 \mathrm{d}x$$

- Specify $u, v$ and BC to ensure $\lambda_\ell > 0$
  - E.g., $v > 0$ and the Dirichlet BC $f(L) = f(R) = 0$

# Positive Eigenvalues

- Integration by parts

$$\lambda_\ell = \langle \mathcal{L}\phi_\ell, \phi_\ell \rangle = - u(x)\phi_\ell(x)\phi_\ell'(x)\Big|_{x=L}^{R} + \int_L^R u(x)[\phi_\ell'(x)]^2 \mathrm{d}x + \int_L^R v(x)[\phi_\ell(x)]^2 \mathrm{d}x$$

- Specify $u, v$ and BC to ensure $\lambda_\ell > 0$
  - E.g., $v > 0$ and the Dirichlet BC $f(L) = f(R) = 0$

**Theorem (Ding and Z, 2017)**

Suppose the S-L equation is regular with $v > 0$ and the Dirichlet BC. Then, its Green's function is an MCF.

## Examples

- Assume $u(x) \equiv 1$, $v(x) \equiv \nu$, and $w(x) \equiv 1$ on $[0, 1]$:

$$-f''(x) + \nu f(x) = 0$$

| | Dirichlet | Cauchy |
|---|---|---|
| $\nu = 0$ | $p(x) = x$<br>$q(y) = 1 - y$ | $p(x) = x$<br>$q(y) = 1$ |
| $\nu > 0$ | $p(x) = \sinh(\gamma x)$<br>$q(y) = \sinh(\gamma(1 - y))$ | $p(x) = d(\gamma)\sinh(\gamma x)$<br>$q(y) = \cosh(\gamma(1 - y))$ |
| $\nu < 0$ | $p(x) = \sin(\gamma x)$<br>$q(y) = \sin(\gamma(1 - y))$ | $p(x) = d(\gamma)\sin(\gamma x)$<br>$q(y) = \cos(\gamma(1 - y))$ |

- $\gamma = \sqrt{|\nu|}$

### Corollary (Ding and Z, 2017)

Let $g(x,y)=\eta[p(x)q(y)\,\mathbb{I}_{\{x\le y\}}+p(y)q(x)\,\mathbb{I}_{\{x>y\}}]$ be as given on the last page. Suppose that $\{x_1,\ldots,x_n\}\subset(0,1)$, where $x_i=x_1+(i-1)h$ with $h=\frac{x_n-x_1}{n-1}$. Then,

$$\boldsymbol{G}^{-1}=\eta^{-1}a\begin{pmatrix} b & -1 & & & \\ -1 & c & -1 & & \\ \cdots & & \cdots & & \cdots \\ & & -1 & c & -1 \\ & & & -1 & d \end{pmatrix}.$$

## Corollary (Ding and Z, 2017)

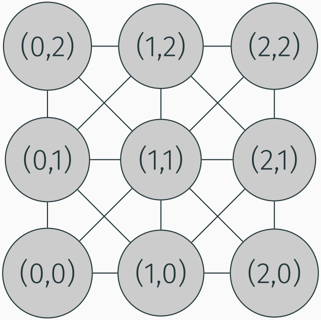
Let $g(x,y) = \eta[p(x)q(y)\,\mathbb{I}_{\{x \le y\}} + p(y)q(x)\,\mathbb{I}_{\{x>y\}}]$ be as given on the last page. Suppose that $\{x_1, \ldots, x_n\} \subset (0,1)$, where $x_i = x_1 + (i-1)h$ with $h = \frac{x_n - x_1}{n-1}$. Then,

$$\boldsymbol{G}^{-1} = \eta^{-1} a \begin{pmatrix} b & -1 & & & \\ -1 & c & -1 & & \\ \cdots & & \cdots & & \cdots \\ & & -1 & c & -1 \\ & & & -1 & d \end{pmatrix}.$$

- $\mathcal{O}(1)$ computation
- Reparameterization

# Extension for $d > 1$

- "Compositional" covariance: $k(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d} k_i(x^i, y^i)$
- Limitation: $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ must form a regular lattice
- Then, $K = \bigotimes_{i=1}^{d} K_i$ and $K^{-1} = \bigotimes_{i=1}^{d} K_i^{-1}$, preserving sparsity



# Numerical Experiments

# Three Covariance Functions

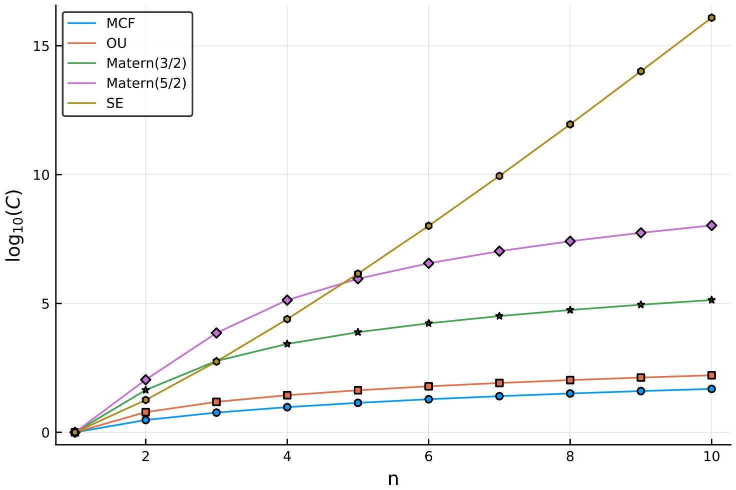
- MCF: $k(x,y) = \eta[p(x)q(y)\,\mathbb{I}_{\{x\leq y\}} + p(y)q(x)\,\mathbb{I}_{\{x>y\}}]$

$$\begin{cases} p(x) = \sin(\sqrt{|\nu|}x), & q(x) = \sin(\sqrt{|\nu|}(1-x)), \quad \text{if } \nu < 0 \\ p(x) = x, & q(x) = 1-x, \quad \text{if } \nu = 0 \\ p(x) = \sinh(\sqrt{\nu}x), & q(x) = \sinh(\sqrt{\nu}(1-x)), \quad \text{if } \nu > 0 \end{cases}$$

- OU: $k(x,y) = \eta\exp(-\theta|x-y|)$
- SE (squared exponential): $k(x,y) = \eta\exp(-\theta(x-y)^2)$

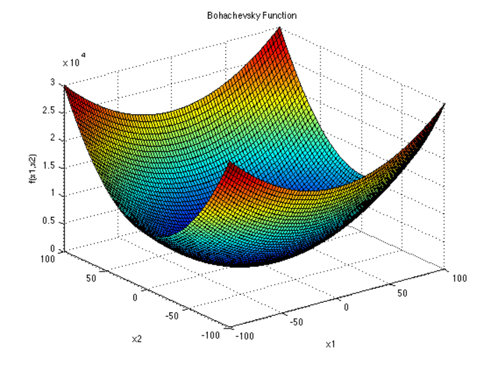
# Condition Number of Covariance Matrix

- $C = \lambda_{\max}(K)/\lambda_{\min}(K)$ measures “closeness to singularity”

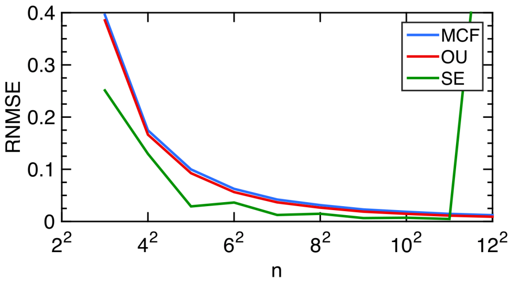


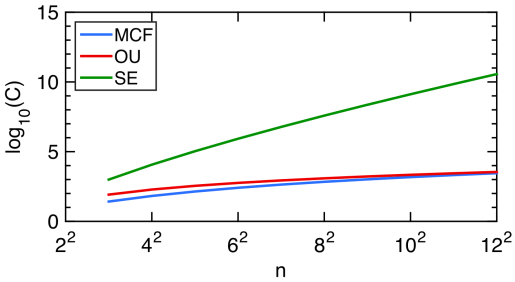
# Artificial Surface: Bohachevsky Function

$$f(x,y) = x^2 + 2y^2 - 0.3\cos(3\pi x) - 0.4\cos(4\pi y) + 0.7$$

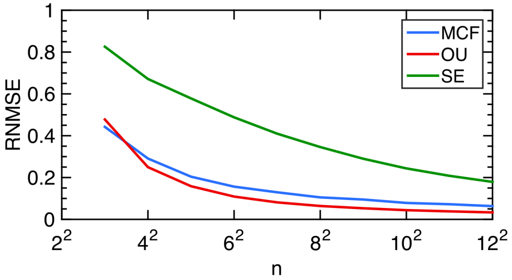


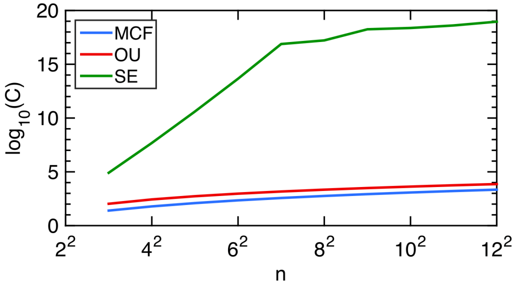
# Prediction with Noise-free Samples



# Prediction with Noisy Samples

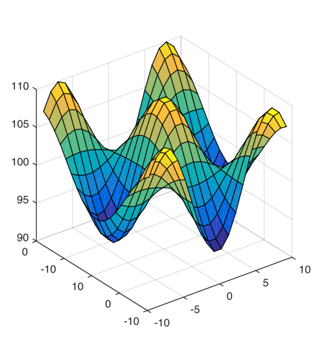


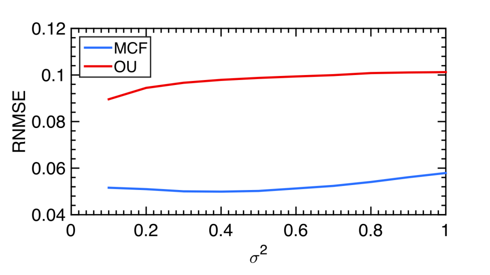



# Artificial Surface: Griewank Function

$$f(\mathbf{x}) = \sum_{i=1}^{d} \frac{x_i^2}{4000} - \prod_{i=1}^{d} \cos(i^{-1/2} x_i) + 1$$

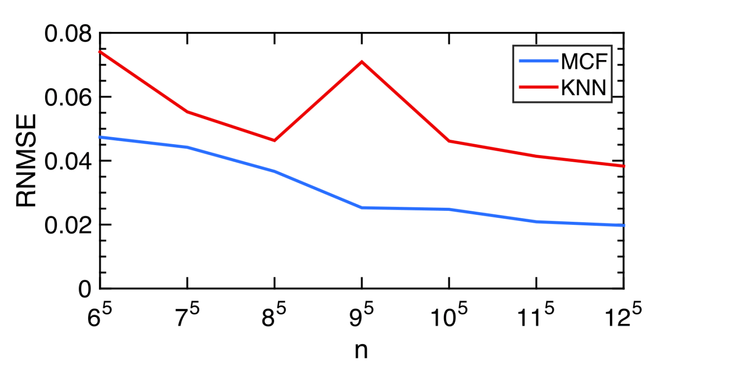
- $d = 5$
- $n = 6^d = 7776$



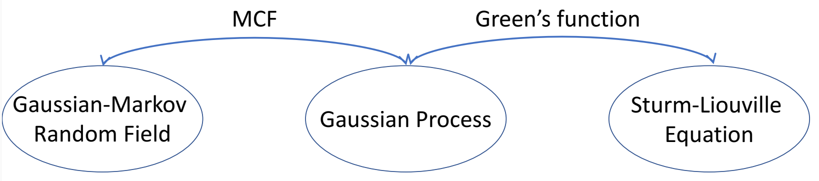
# Artificial Surface: Sphere Function

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^2$$



# Conclusions

# Conclusions




- MCFs allow modeling association directly, while retaining sparsity in the precision matrix
- Reduce computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$ without approximations
  - Further reduce to $\mathcal{O}(n)$ if observations are noise-free
- Enhance numerical stability substantially
- Limitation: design points must form a regular lattice, though not necessarily equally spaced

Markovian covariances without approx.

v.s.

Good approx. for *all* covariances