Priors for Covariance Operators

Dennis D. Cox
Rice University

with Hongxiao Zhu

Introduction

- **Objective:** develop priors for Bayesian functional data analysis.
- Assume data area realizations of a Gaussian process, ...
- ... say we observe $Y_i(t)$, $t \in [a, b]$ where
- Y_1, Y_2, $\ldots Y_n$ are i.i.d. $N(\mu, V)$:

$$\mu(t) = E[Y(t)], \quad V(t, s) = \text{Cov}[Y(t), Y(s)].$$

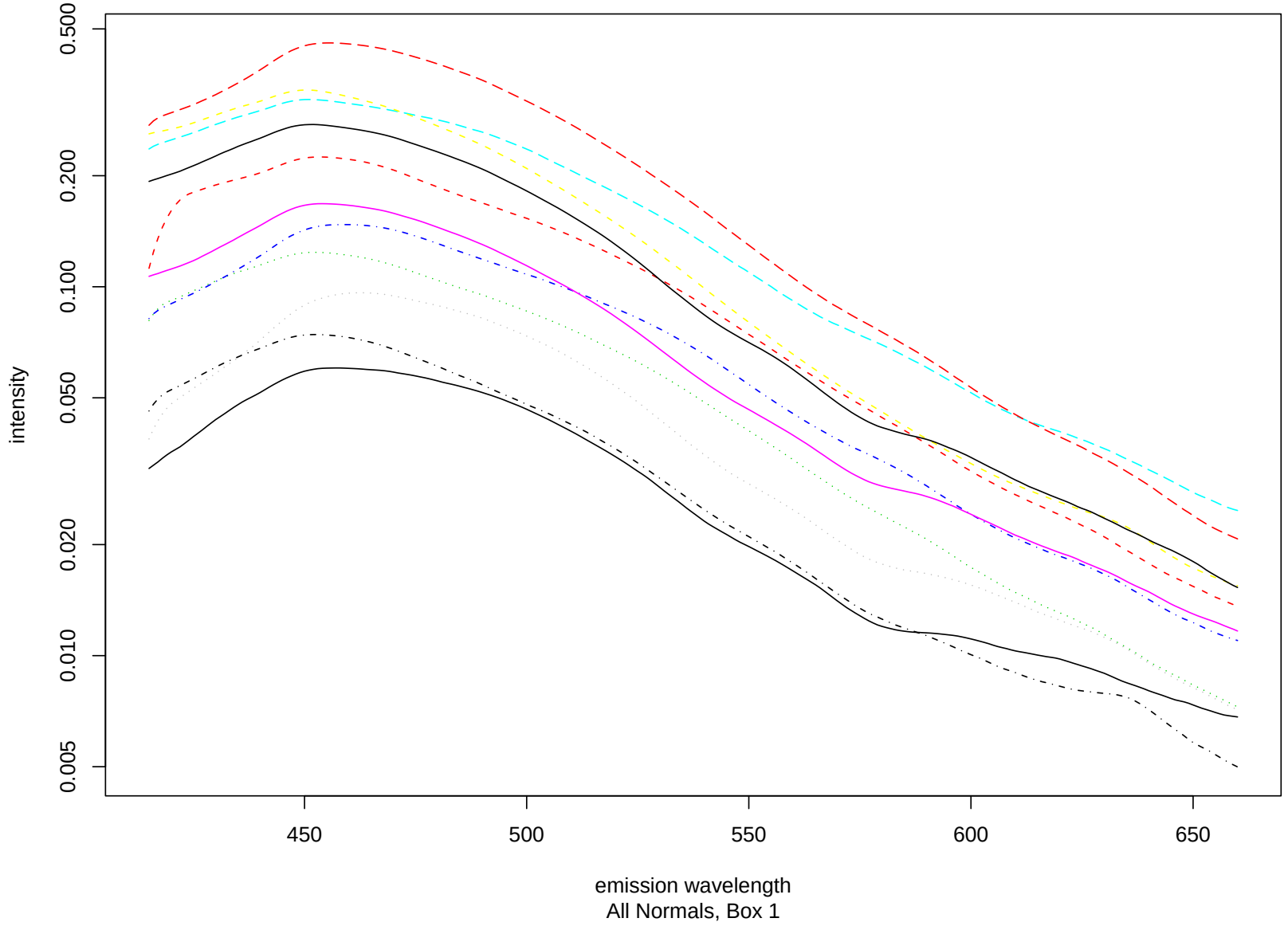
- $\mu | V, k \sim N(0, kV)$, $k \sim \text{InverseGamma}(\alpha, \beta)$,
- But $V \sim$?????

Outline of Presentation

1. Functional Data
2. Requirements on covariance priors
3. Some things that don't work
4. A proposal
5. Application to simulated data
6. Topics for further research

Functional Data:

- "Functional data" means the data $Y_i(\cdot)$ are functions of a continuous variable
- Some books on the subject: *Functional Data Analysis* and *Applied Functional Data Analysis* by Ramsay and Silverman; *Nonparametric Functional Data Analysis: Theory and Practice* by Ferraty and Vieu.
- We assume here that the data have been smoothed and registered.
- Example: Fluorescence spectroscopy measurements


Ten Selected Spectra at Ex=360nm
intensity
0.500
0.200
0.100
0.050
0.020
0.010
0.005
450
500
550
600
650
emission wavelength
All Normals, Box 1

Functional Data (cont.):

- Working with functional data requires some idealization
- E.g. the data are actually multivariate; they are stored as either of

  (G) $(Y_i(t_1), \ldots, Y_i(t_m))$, vectors of values on a grid

  (C) $(\eta_{i1}, \ldots, \eta_{im})$ where $Y_i(t) = \sum_j \eta_{ij} B_j(t)$ is a basis function expansion (e.g., B-splines).
- Note that the order of approximation m is rather arbitrary.
- Treating functional data as simply multivariate doesn't make use of the additional "structure" implied by being a smooth function.

Functional Data (cont.):

- Methods for Functional Data Analysis (FDA) should satisfy the *Grid Refinement Invariance Principle (GRIP)*:
- as the order of approximation becomes more exact (i.e., $m \to \infty$), the method should approach the appropriate limiting analogue for true functional (infinite dimensional) observations.
- Denote the discretized data by $\vec{Y}_i^{(m)} = \vec{Y}_i = (Y_i(t_1), \ldots, Y_i(t_m))$ and the corresponding mean vectors and covariance matrix $\vec{\mu}$ and $\vec{V}$ where $\vec{V}_{ij} = V(t_i, t_j)$.

Functional Data (cont.):

- The GRIP rules out use of a pivotal quantity based on Hotellings T^2 for constructing confidence sets for μ:
- $T^2(\mu) = (\vec{\bar{Y}} - \vec{\mu})^T \hat{\vec{\vec{V}}}^{-1} (\vec{\bar{Y}} - \vec{\mu})$
- When $m >= n$, $\hat{\vec{\vec{V}}}$ is not invertible.

Requirements on Covariance Priors:

- Our first requirement in constructing a prior for covariance functions is that we mind the GRIP
- For example, we can't simply use $\vec{V}^{-1} \sim Wishart(d_m, W_m)$ for some $m \times m$ matrix W_m.

Requirements on Covariance Priors (cont.):

- Perhaps the easiest way to satisfy the GRIP requirement is to construct a prior on the space of covariance functions and then project it down to the finite dimensional approximation.
- For example, using grid values, $\vec{V}_{ij} = V(t_i, t_j)$.
- For this purpose, it is useful to think of the Y as a random variable taking values in a real separable Hilbert space H.
- The space of covariance operators S for Gaussian measures on H is the set of self adjoint nonnegative definite trace class operators:

Requirements on Covariance Priors (cont.):

- It will be useful to define the space of trace class operators and the corresponding norm.
- First, the adjoint of a linear operator $A : H \longrightarrow H$ is defined by $\langle x, Ay \rangle = \langle A^* x, y \rangle$.
- A bounded linear operator $A : H \longrightarrow H$ is *trace class* iff for any othonormal basis $e_1, e_2, \ldots$

$$\|A\|_{\mathcal{L}_1} = \sum_i \langle (A^* A) e_i, e_i \rangle^{1/2} < \infty$$

- The $\mathcal{L}_1$ norm is stronger than the usual operator norm given by $\|A\| = \sup\{\|Au\| : \|u\| = 1\}$.

Requirements on Covariance Priors (cont.):

- The space S of covariance operators satisfies $A \in S$ iff
  1. A is bounded linear operator $A : H \longrightarrow H$ which is self-adjoint, i.e. $A^* = A$.
  2. A is nonnegative definite, i.e. $\langle x, Ax \rangle_H \geq 0$ for all $x \in H$.
  3. A is a trace class operator.
- One can show that $A \in S$ is equivalent to
  (a) There is a complete orthormal basis $\phi_1, \phi_2, \ldots$ of eigenvectors and eigenvalues $\lambda_i \geq 0$ such that $A = \sum_i \lambda_i \phi_i \otimes \phi_i$ where $\otimes$ is defined by $(x \otimes y)z = \langle x, z \rangle y$.
  (b) The trace class condition: $\sum_i \lambda_i < \infty$.

Requirements on Covariance Priors (cont.):

- The next requirement is that the prior be *fully* nonparametric
- An example that would not be fully nonparametric is a prior on stationary covariance operators.
- A covariance function $V(s,t)$ is stationary iff $V(s,t) = v(s-t)$ for some function v of one variable.
- It is relatively easy to put a prior on the stationary covariances. For example, by Bochner's theorem $v(t) = \int \exp[i\omega t] d\mu(\omega)$ for some symmetric nonnegative measure μ (i.e., $v(t)/v(0)$ is the characteristic function of a symmetric probability measure). One could use a Dirichlet process to construct an appropriate prior on μ.

Requirements on Covariance Priors (cont.):

- We will take "fully nonparametric" to mean that the support of the prior must be all of S.
- The support of a probability measure is the smallest closed subset of probability 1. Alternatively, it is the set of points whose neighborhoods all have positive probability.
- The class of stationary covariances is a proper closed subset of S. So, it is not sufficient to only consider the stationary covariances.

Requirements on Covariance Priors (cont.):

- Our final requirement is that the prior admit a computable posterior.
- In summary, we want
  1. a prior on S
  2. a prior on all of S
  3. a prior we can compute with

Some things that don't work:

- It would be very nice if we could construct a conjugate prior like the inverse Wishart in finite dimensions.
- This doesn't seem feasible. The main difficulty is that the inverse of a covariance operator is not bounded.
- For example, let $Y(t)$ be Brownian motion considered as taking values in $L_2[0,1]$. Then $v(s,t) = \mathrm{Cov}(Y(t), Y(s)) = \min\{s,t\}$.
- The operator V is defined by

$$Vf(s) = \int_0^1 v(s,t)f(t)dt.$$

- Try to compute $V^{-1}g$ by solving the integral equation

$$g(s) = \int_0^1 v(s,t)f(t)dt$$

for f.

Some things that don't work (cont.):

- With a little calculus

$$\begin{aligned} g(s) &= \int_0^1 \min(s,t) f(t) dt \\ &= \int_0^s t f(t) dt + s \int_s^1 f(t) dt. \end{aligned}$$

- We see g is absolutely continuous and $g(0) = 0$. Differentiating

$$\begin{aligned} g'(s) &= s f(s) - s f(s) + \int_s^1 f(t) dt \\ &= \int_s^1 f(t) dt \end{aligned}$$

- We see g' is absolutely continuous and $g'(1) = 0$. Differentiating again

$$g''(s) = -f(s).$$

Some things that don't work (cont.):

- Thus, in the Brownian motion case, V is invertible at g iff g' is absolutely continuous and satisfies the two boundary conditions. Thus, V is certainly not invertible on all of $L^2[0,1]$.
- We can understand the problem in general by going back to the spectral representation:

$$V = \sum_i \lambda_i \phi_i \otimes \phi_i.$$

- Thus $Vx = \sum_i \lambda_i \langle x, \phi_i \rangle \phi_i$
- Then, if $V^{-1}x$ exists, it is given by

$$V^{-1}x = \sum_i \lambda_i^{-1} \langle x, \phi_i \rangle \phi_i$$

- This converges in H iff $\sum_i \lambda_i^{-2} \langle x, \phi_i \rangle^2 < \infty$, which is a pretty strict condition on x since $\sum_i \lambda_i < \infty$.

Some things that don't work (cont.):

- We say V^{-1} is an unbounded operator in H, and we have to specify its domain. This domain is generally pretty complicated (e.g. for most common covariance operators, differentiability conditions and boundary conditions).
- Thus, putting a prior on V^{-1} looks to be very problematic.
- What about putting a prior on $\log V$? (Chiu, Leanord, Tsui: "The Matrix-Logarithmic Covariance Model" *JASA* 1996).

$$\log V = \sum_i (\log \lambda_i) \phi_i \otimes \phi_i.$$

This is also an unbounded operator in H (note that $\log \lambda_i \to -\infty$).

A proposed approach that does work:

- Suppose Z_1, Z_2, ... are i.i.d. H-valued random variables with $N(0, B)$ distribution.
- Consider

$$V = \sum_i w_i Z_i \otimes Z_i$$

  where w_1, w_2, ... are nonnegative constants.
- **Theorem 1:** If $\sum_i w_i < \infty$, then the series defining V converges a.s. in $\mathcal{L}_1$ norm. Thus, V is a random variable taking values in S.
- **Theorem 2:** If in addition $B > 0$ and $w_i > 0$ for all i, then the support of the distribution of V is all of S.

A proposed approach that does work (cont.):

- Thus, if we can compute with this proposed prior, we will have satisfied all three requirements.
- Assuming we use values on a grid for the finite dimensional representation, let $\vec{Z}_i = (Z(t_1), \ldots, Z(t_m))$. Then

$$\vec{V} = \sum_i w_i \vec{Z}_i \vec{Z}_i^T$$

- How to compute with this? One idea is to write out the characteristic function and use Fourier inversion. That works well for weighted sum of χ^2 distributions (fortran code available from Statlib)

A proposed approach that does work (cont.):

- Another approach: use the $\vec{Z}_i$ directly. We will further approximate $\vec{V}$ by truncating the series:

$$\vec{V}^{(m,j)} = \sum_{i=1}^{j} w_i \vec{Z}_i \vec{Z}_i^T$$

Let

$$\boldsymbol{Z}^{(m,j)} = \begin{bmatrix} w_1^{1/2} \vec{Z}_1^T \\ w_2^{1/2} \vec{Z}_2^T \\ \vdots \\ w_j^{1/2} \vec{Z}_j^T \end{bmatrix}$$

then

$$\vec{V} = \boldsymbol{Z}^T \boldsymbol{Z}$$

A proposed approach that does work (cont.):

- Writing $\boldsymbol{Z}$ in terms of its Q-R decomposition:

$$\boldsymbol{Z = QR},$$

where $\boldsymbol{Q^T Q = I}$ and $\boldsymbol{R}$ is upper triangular with nonnegative diagonal, we have

$$\vec{\boldsymbol{V}} = \boldsymbol{R^T Q^T QR = R^T R},$$

so $\boldsymbol{R}$ is the Cholesky factor for $\vec{\boldsymbol{V}}$. We will need $\vec{\boldsymbol{V}}^{-1}$ and $\det \vec{\boldsymbol{V}}$ to compute the likelihood, and the most convenient way to obtain these is through the Cholesky factorization.

- But most importantly, there are efficient algorithms for updating the QR decomposition of a matrix $\boldsymbol{A}$ when one rows or columns of $\boldsymbol{A}$ are changed.

A proposed approach that does work (cont.):

Here is the inner loop of a Metropolis-Hastings type algorithm

- For $i = 1, j$ {
  - Random walk proposal:

$$\vec{Z}_i^p = \vec{Z}_i + \nu_i$$

  - R^p, the Cholesky factor for $\vec{V}^p$ is obtained by rank one QR update
  - Compute the unnormalized posterior

$$(\det R^p)^{-(n+1/2)} \exp[-.5 \sum_k \|(R^p)^{-1/2}(\vec{Y}_k - \vec{\mu}_k)\|^2$$

$$-.5c^{-1}\|(R^p)^{-1/2}\vec{\mu}_k\|^2] \exp[-.5(Z_i^p)^T \vec{B}^{-1}(Z_i^p)]$$

A proposed approach that does work (cont.):

- Compute the Metropolis ratio and accept or reject the proposed Z_i^p.
- }

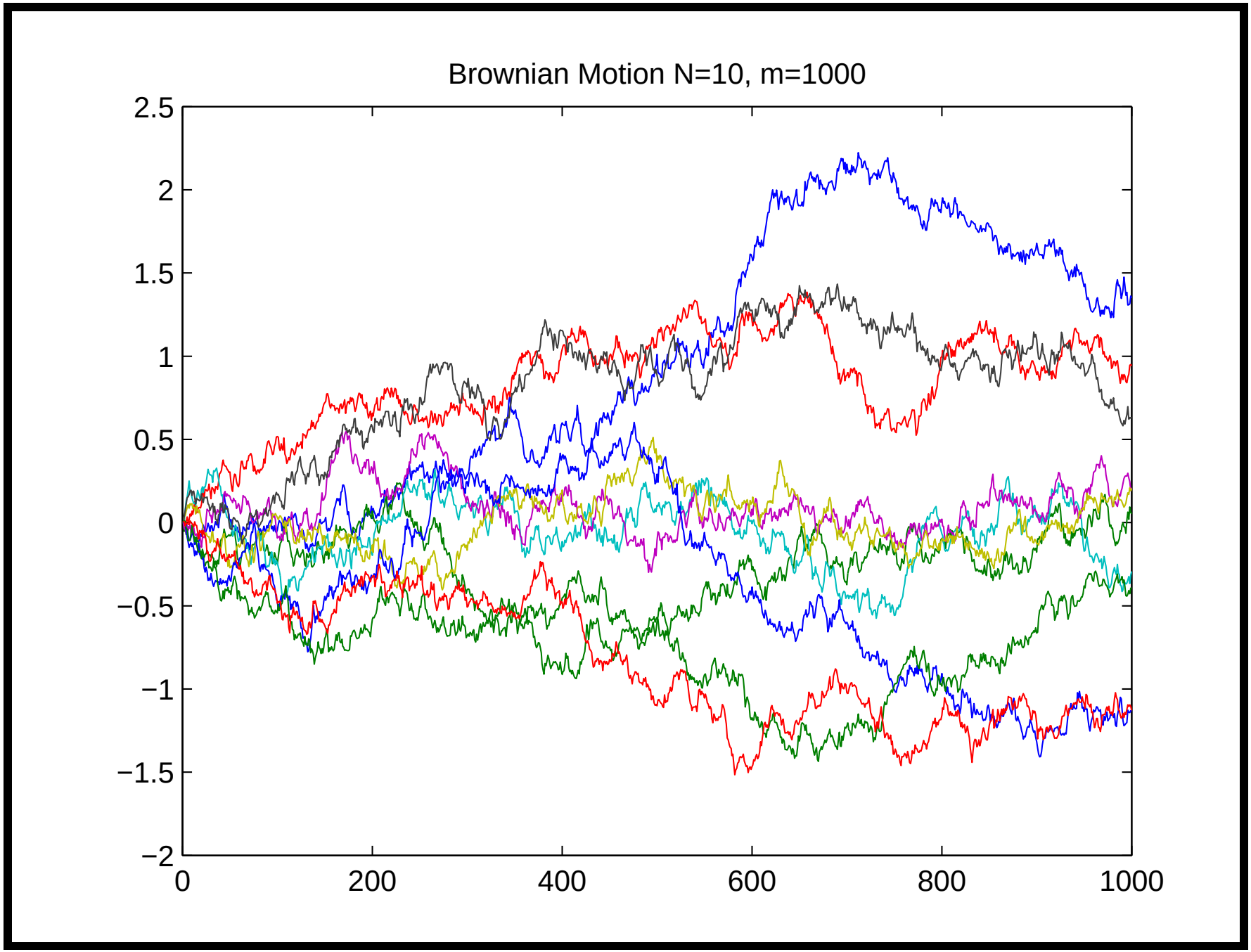
Update $\vec{\mu}$ by generating from $N((n + c^{-1})^{-1}\bar{\vec{Y}}, (n + c^{-1})^{-1}\vec{V})$

A proposed approach that does work (cont.):

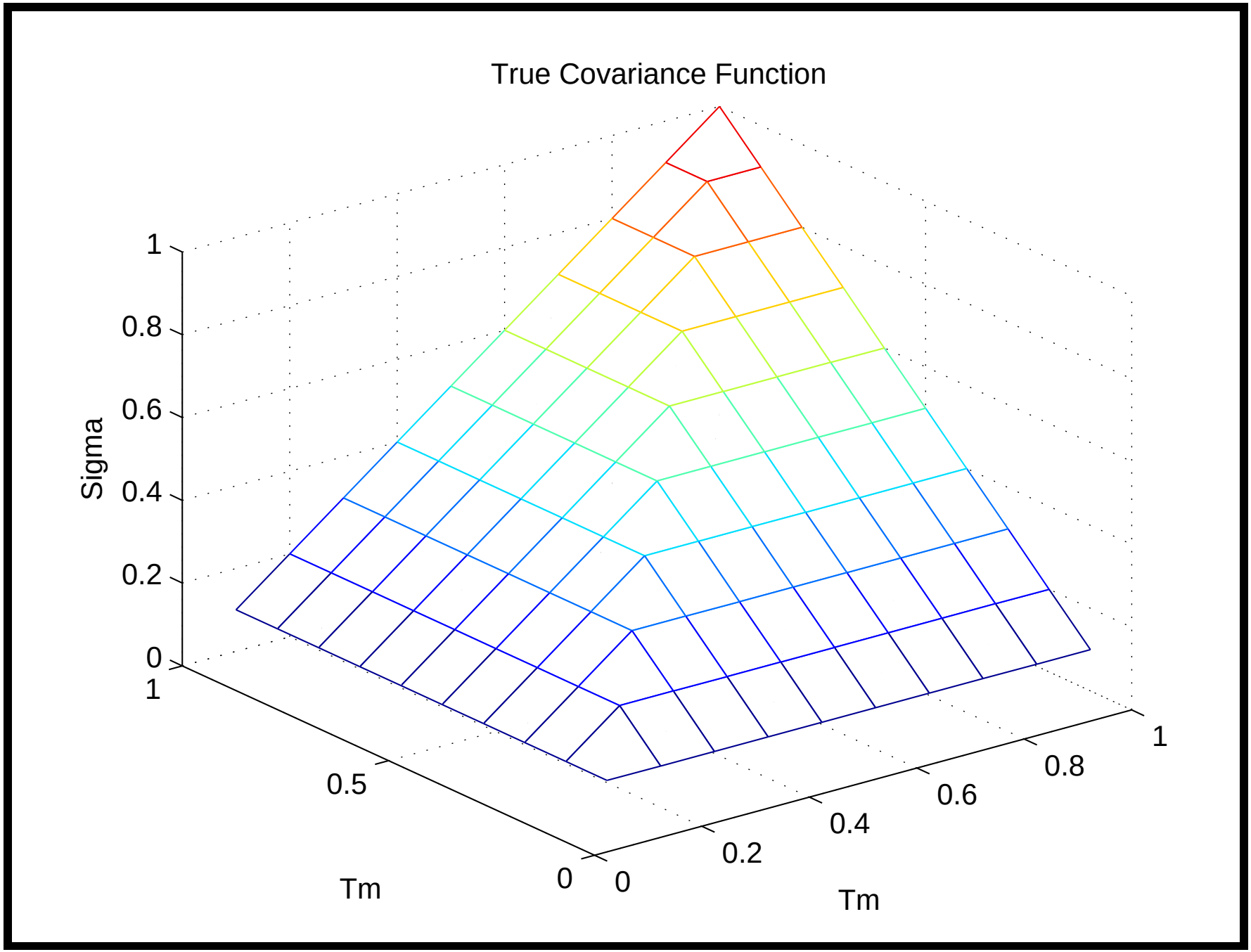
- There are a couple of minor modifications:
  1. We include an additional scale parameter k in $V = k \sum_i w_i Z_i \otimes Z_i$ where k has an independent inverse Γ prior.
  2. We actually use the marginal unnormalized posterior $f(\boldsymbol{Z}|\vec{Y}_1, \ldots \vec{Y}_n)$ in the updating, i.e. we integrate out μ.
- The algorithm has been implemented in Matlab.

Some results with simulated data:

- First experiment: Generated data from Brownian motion (easy to do!)
- $n = 50$ and various values of m and j


Brownian Motion N=10, m=1000
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
200
400
600
800
1000

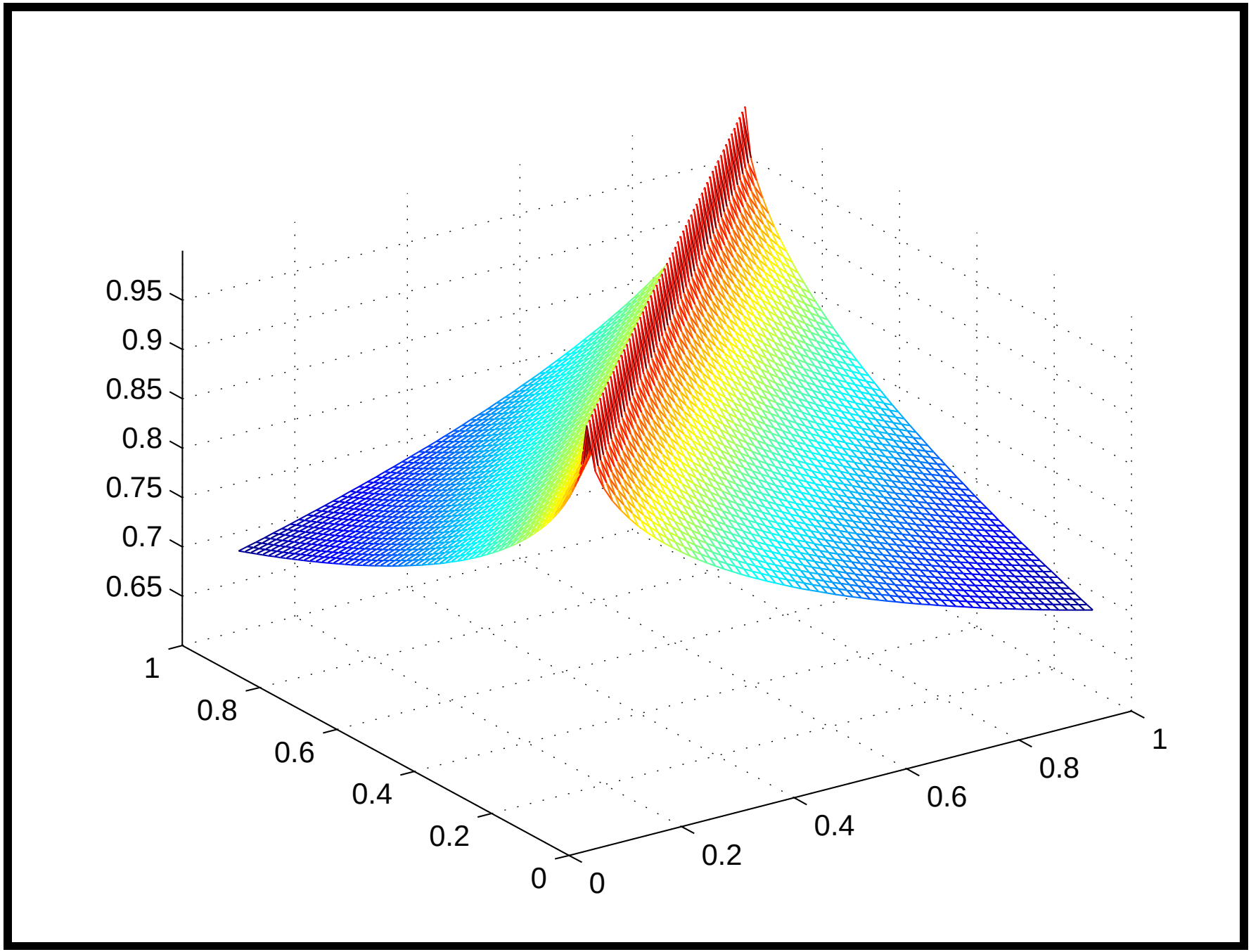
First, the True Covariance function for Brownian Motion.


True Covariance Function
Sigma
1
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

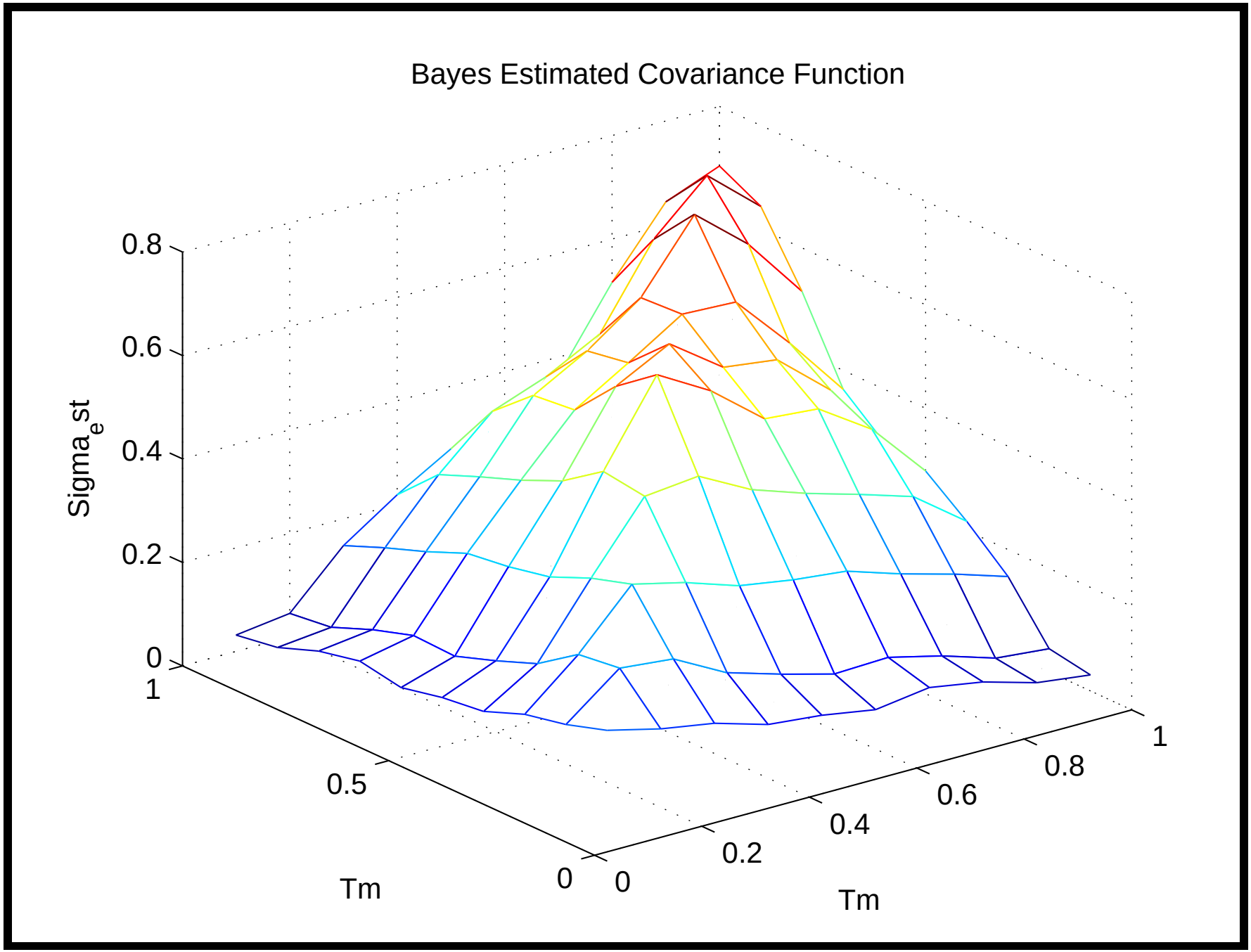
The covariance function used to generate the Z_i is the Ornstein-Uhlenbeck correlation:

$$B(s,t) = \exp[-\alpha|s-t|]$$

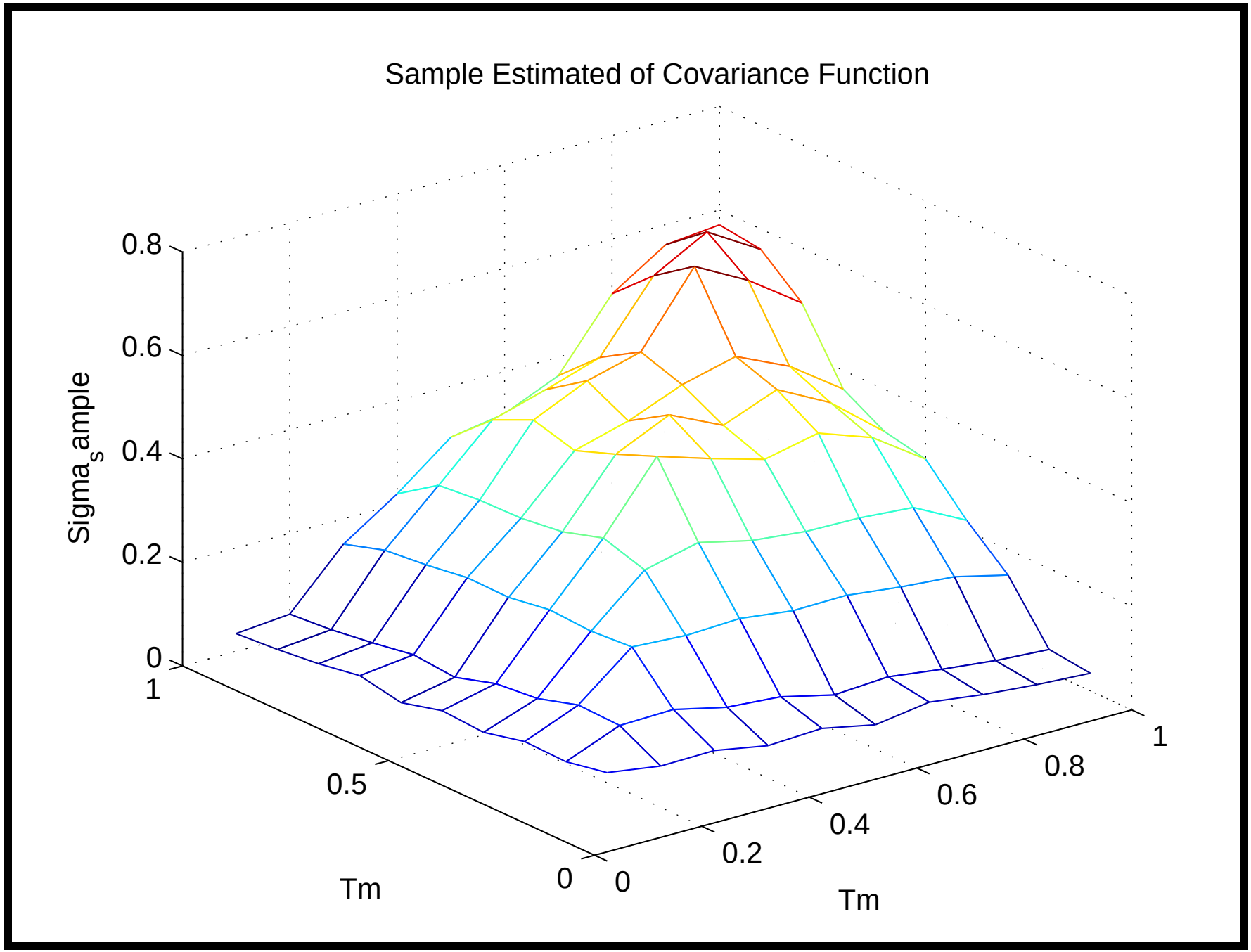
with $\alpha = 1$. This process goes by a number of other names (the Gauss-Markov process, Continuous Time Autoregression of order 1, etc.)


0.95
0.9
0.85
0.8
0.75
0.7
0.65
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

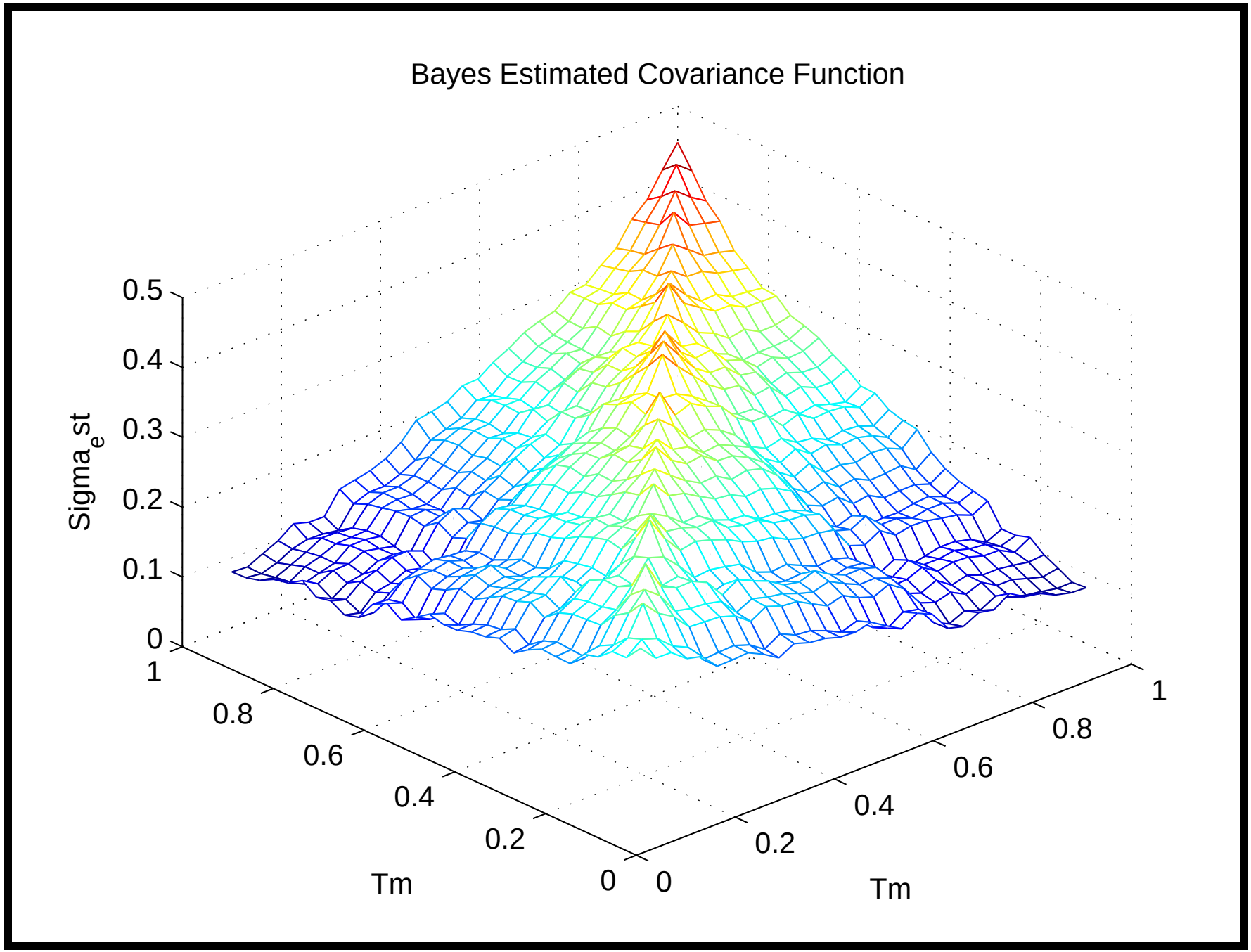
The Bayesian posterior mean estimate with $m = 10$, $j = 20$.


Bayes Estimated Covariance Function
Sigma_est
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

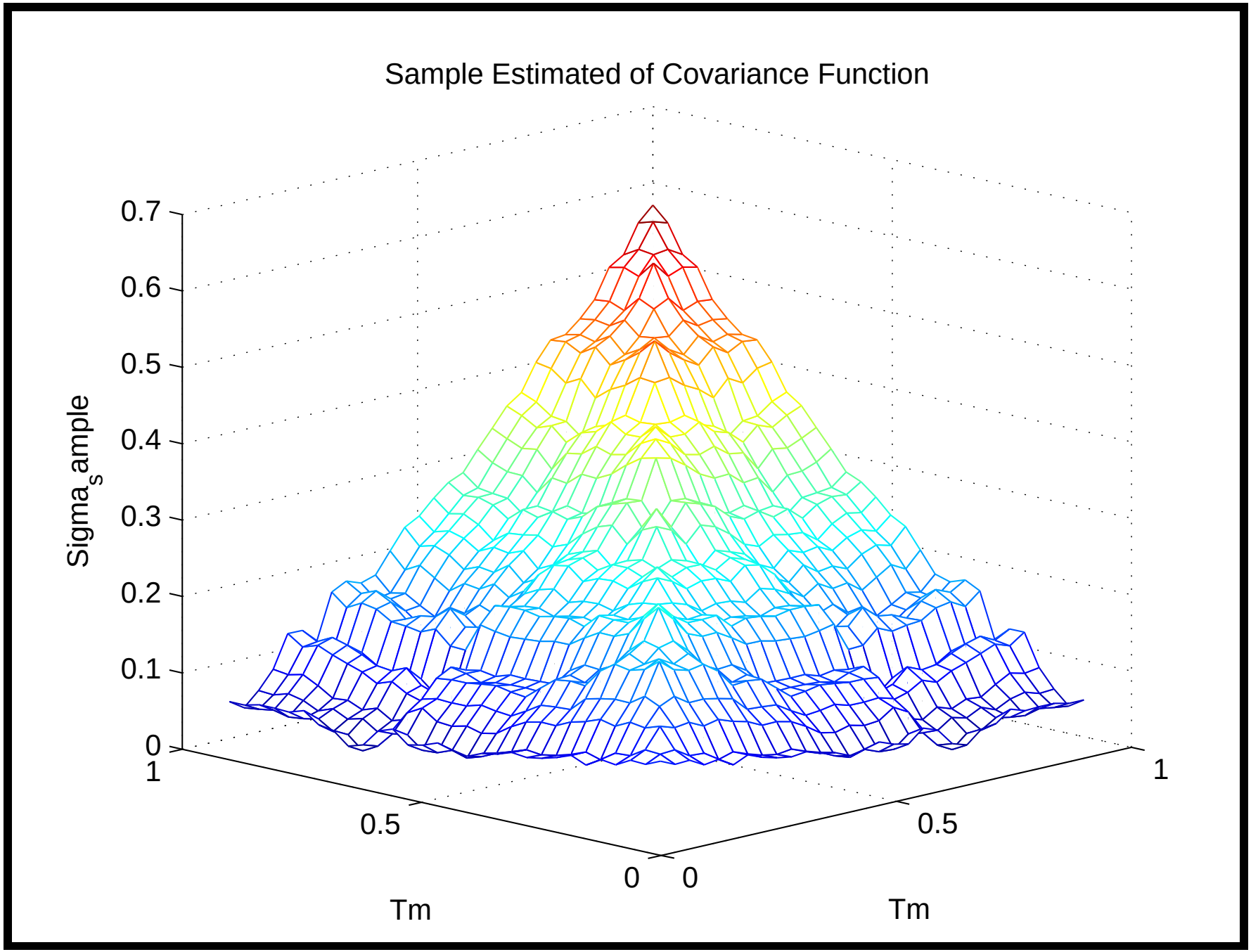
The sample covariance estimate with $m = 10$.


Sample Estimated of Covariance Function
Sigma_sample
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

Now the Bayes posterior mean estimate with $m = 30$, $j = 60$.


Bayes Estimated Covariance Function
Sigma_est
0.5
0.4
0.3
0.2
0.1
0
1
0.8
0.6
0.4
0.2
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

The sample covariance estimate with $m = 30$.


Sample Estimated of Covariance Function
Sigma_sample
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.5
0
Tm
0
0.5
1
Tm

Some results with simulated data:

- Mean squared error results (averaged over the grid points):

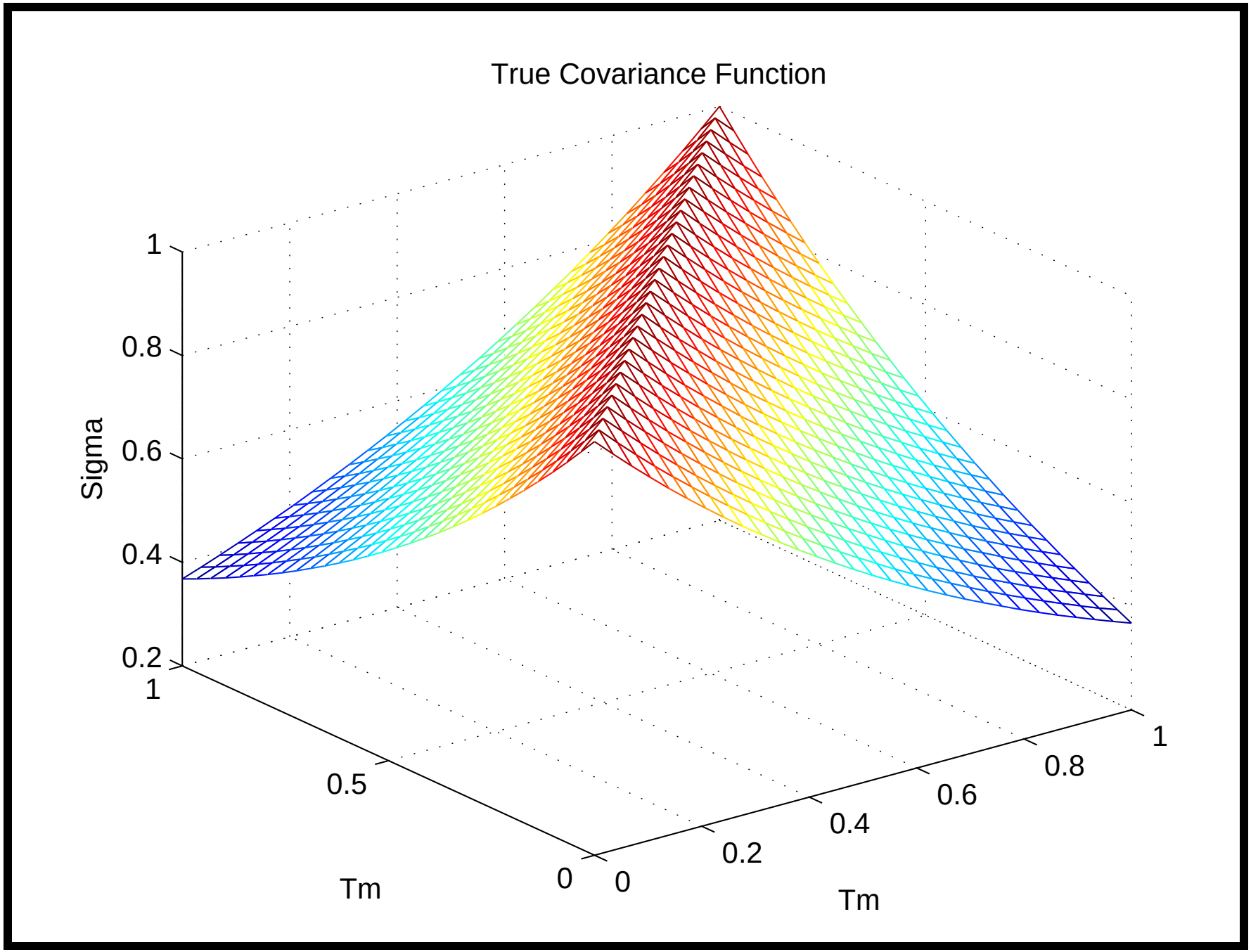
m	j	MSE Bayes	MSE Sample
10	20	0.017	0.026
30	60	0.065	0.054

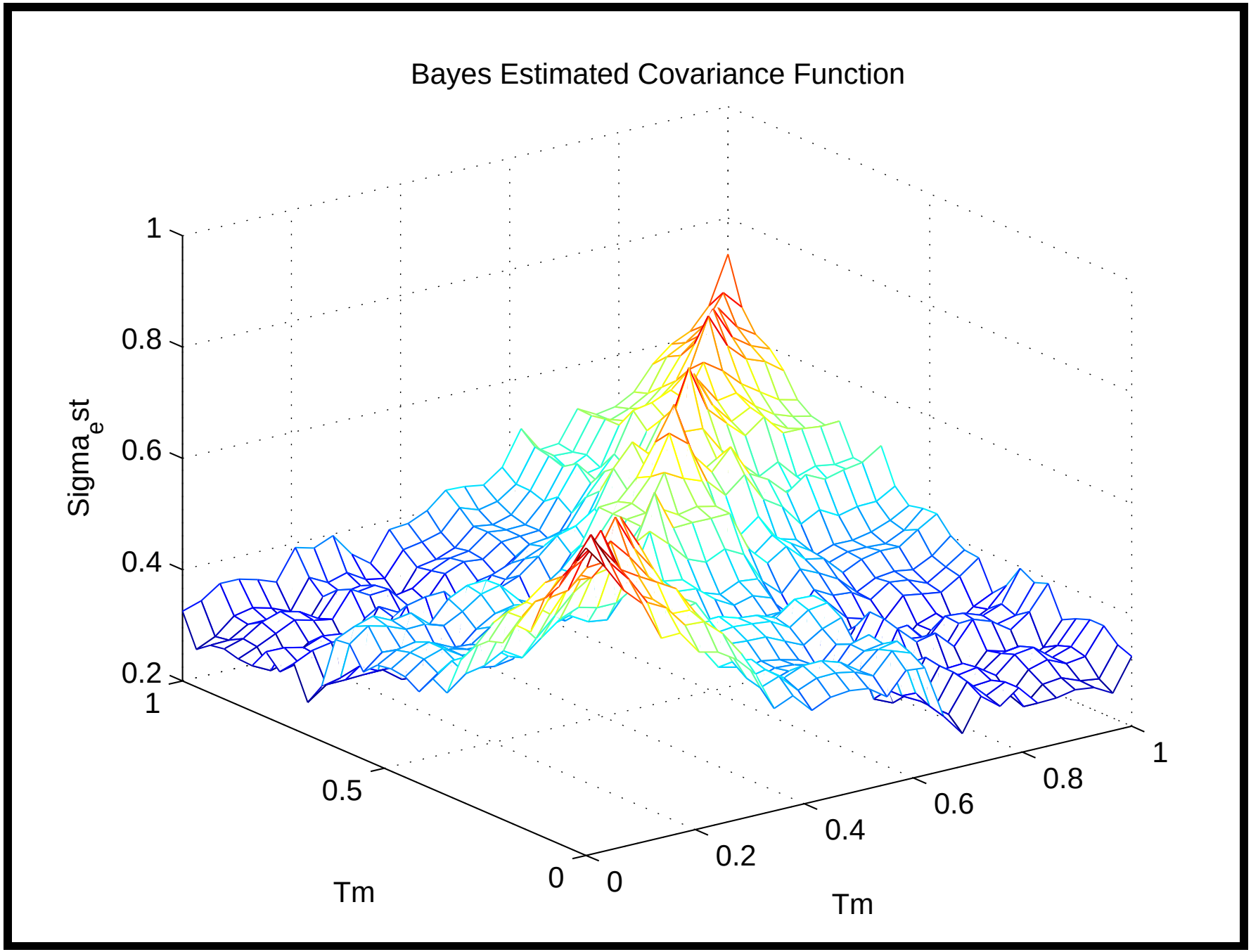
- Next example: we used the Ornstein-Uhlenbeck process with covariance

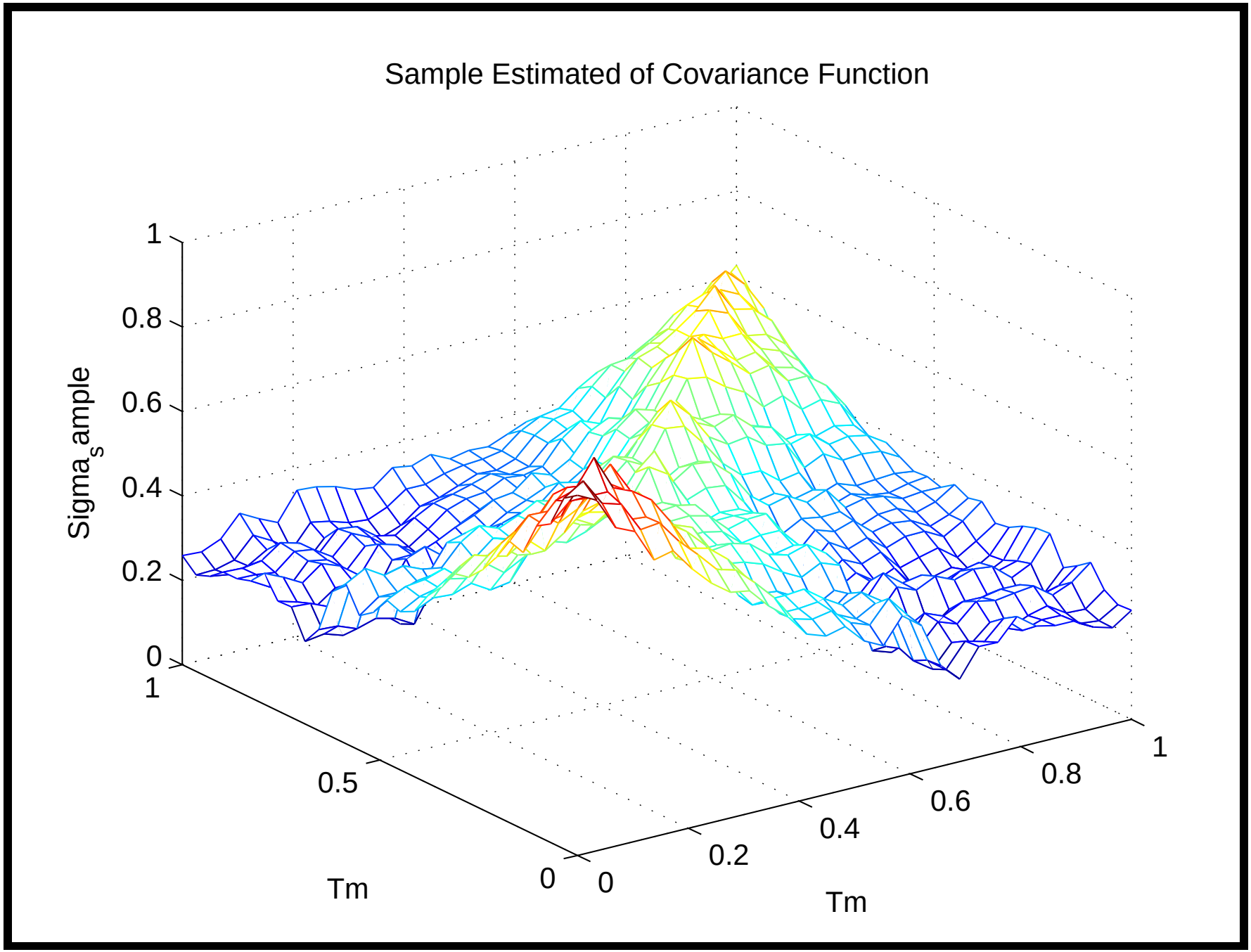
$$\exp[-2|s-t|]$$

  so the prior mean is similar in shape to the true covariance.

- We truth, the Bayes posterior mean, and the sample covariance
- We used $m = 30$ and $j = 60$.


True Covariance Function
Sigma
1
0.8
0.6
0.4
0.2
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm


Bayes Estimated Covariance Function
Sigma_est
1
0.8
0.6
0.4
0.2
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm


Sample Estimated of Covariance Function
Sigma_sample
1
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

Some results with simulated data:

- Mean squared error results are 0.099 and 0.103 for Bayes and sample covariance, respectively

Further research:

- For this first proposal, there are a lot of things to be investigated
- The use of other prior means (than the OU). In particular, I would like to investigate the (partially improper) multiply integrated Brownian motion covariances associated with smoothing splines. We are not seeing much smoothing in the Bayesian posterior mean (but this may be from Monte Carlo uncertainty).
- How to select j, the truncation point in the series $V = \sum_i w_i Z_i \otimes Z_i$.
- How to select the w_i? We have been using $w_i \propto i^{-2}$.

Further research (cont.):

- Are there better ways to structure the computation for speed and accuracy? Computations currently done on a supercomputer with 400 nodes and take hours for the $m = 30$ discretizations.
- Choosing the step distribution in the random walk proposal. We are currently using $N(0, bI)$ for each Z_i and selecting b to get a decent acceptance rate ($\approx 25\%$). Is it better to let b vary with i? Is there a better step distribution?

Further research (cont.):

- There are other possible approaches that might pan out.
- Is there some distribution on S whose finite dimensional projections are inverse Wishart?
- ... a tractable mixture of inverse Wisharts?
- Or can a mixture of inverse Wisharts be used to approximate the finite dimensional projections of a given prior on S?
- Can we construct a prior directly through principal components (spectral decomposition)? I have some ideas but they are very complicated.
- We could consider a Gaussian prior on $\mathcal{L}_2$ (the space of Hilbert-Schmidt operators) giving (say) Z and take $V = Z^*Z$. (Or just self adjoint Z and $V = Z^2$.) This looks computationally feasible with an algorithm very similar to the

one already presented.