Fundamental Issues in Bayesian Functional Data Analysis

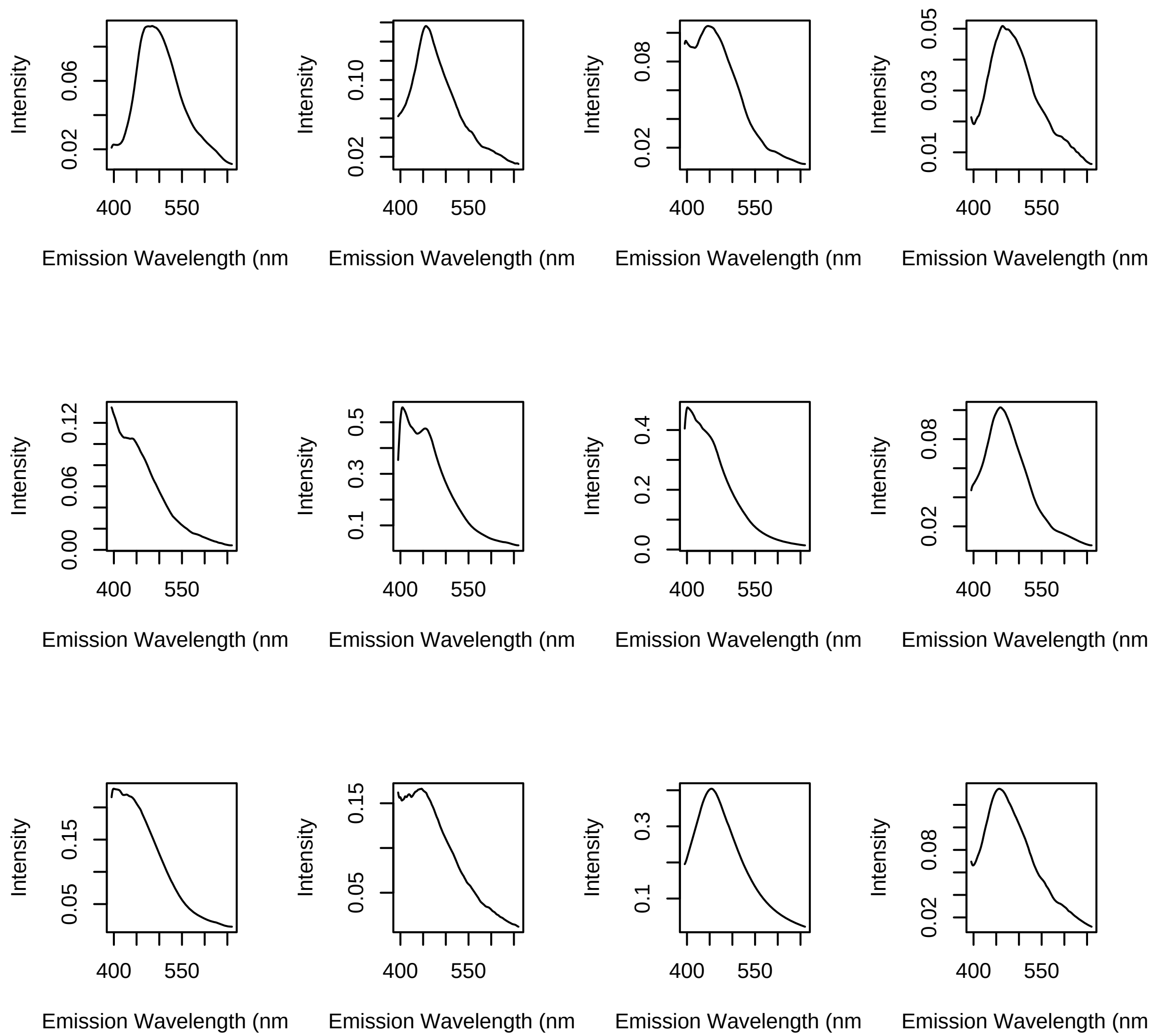
Dennis D. Cox
Rice University

Introduction

- **Question:** *What are functional data?*
- **Answer:** Data that are functions of a continuous variable.
- ... say we observe $Y_i(t)$, $t \in [a, b]$ where
- Y_1, Y_2, $\ldots Y_n$ are i.i.d. $N(\mu, V)$:

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

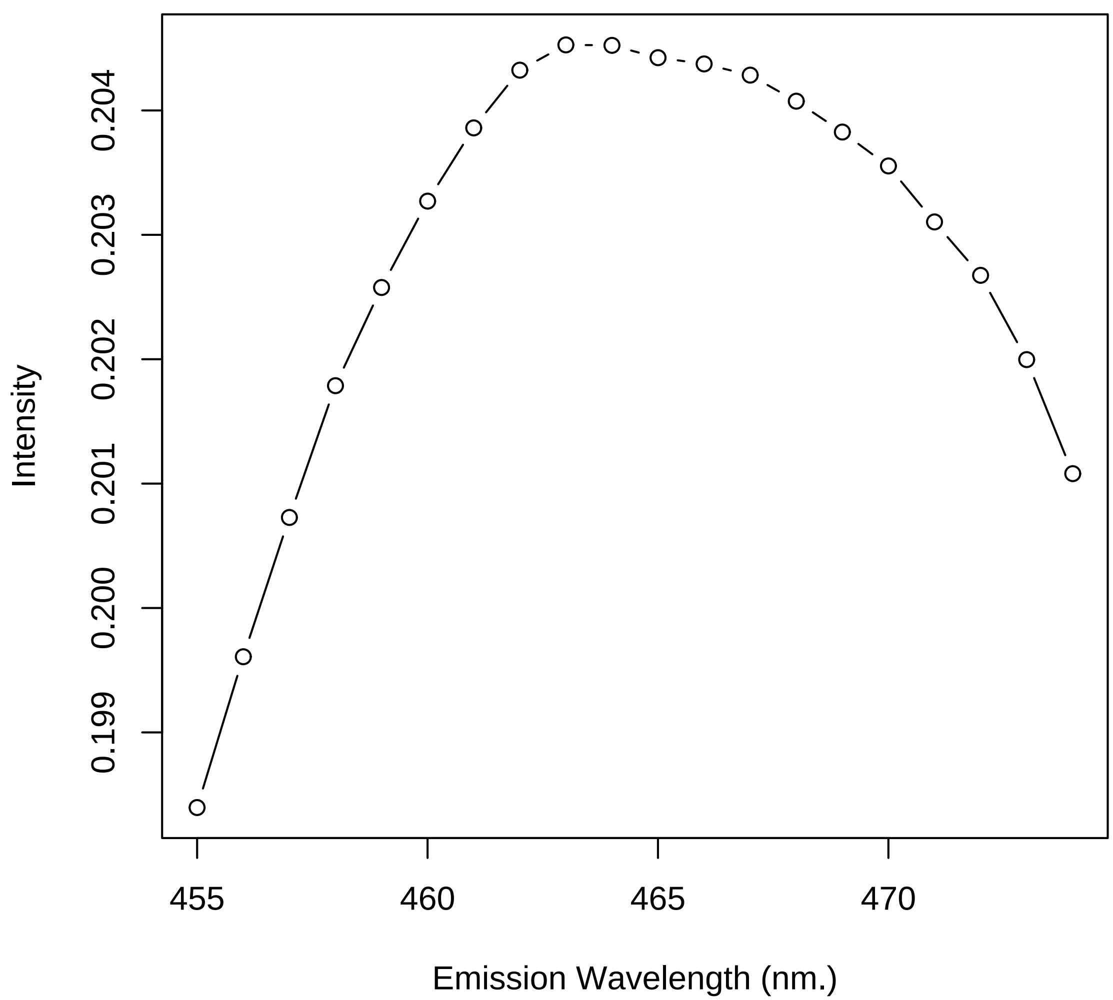
- **Question:** *Do we ever really observe functional data?*
- Here's some examples of functional data:



Intensity
0.02 0.06
400 550
Emission Wavelength (nm
Intensity
0.02 0.10
400 550
Emission Wavelength (nm
Intensity
0.02 0.08
400 550
Emission Wavelength (nm
Intensity
0.01 0.03 0.05
400 550
Emission Wavelength (nm
Intensity
0.00 0.06 0.12
400 550
Emission Wavelength (nm
Intensity
0.1 0.3 0.5
400 550
Emission Wavelength (nm
Intensity
0.0 0.2 0.4
400 550
Emission Wavelength (nm
Intensity
0.02 0.08
400 550
Emission Wavelength (nm
Intensity
0.05 0.15
400 550
Emission Wavelength (nm
Intensity
0.05 0.15
400 550
Emission Wavelength (nm
Intensity
0.1 0.3
400 550
Emission Wavelength (nm
Intensity
0.02 0.08
400 550
Emission Wavelength (nm

Introduction (cont.)

- **Question:** *But you don't really observe continuous functions, do you?*
- **Answer:** Look closely at the data ...



Intensity
0.199
0.200
0.201
0.202
0.203
0.204
455
460
465
470
Emission Wavelength (nm.)

Introduction (cont.)

- OK, so it is really a bunch of dots connected by line segments.
- That is, we really have the data $Y_i(t)$ for t on a grid: $t \in \{395, 396, \ldots, 660\}$.
- But people doing functional data analysis like to pretend they are observing whole functions.
- Is it just a way of sounding erudite? "Functional Data Analysis, not for the heathen and unclean."
- 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.

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 $Y_i(t) = \sum_{j=1}^{m} \eta_{ij} B_j(t)$ 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.*
- Thus the statistical procedure will not be strongly dependent on the finite dimensional approximation.
- Two general ways to mind the GRIP:

  (i) Direct: Devise a method for true functional data, then find a finite dimensional approximation ("projection").

  (ii) Indirect: Devise a method for the finite dimensional data, then see if it has a limit as $m \to \infty$.

See Lee & Cox, “Pointwise Testing with Functional Data Using the Westfall-Young Randomization Method,” *Biometrika* (2008) for a frequentist nonparametric approach to some testing problems with functional data.

Bayesian Functional Data Analysis:

- **Why Bayesian?**
- After all, Bayesian methods have a high "information reqirement," i.e. a likelihood and a prior.
- In principle, statistical inference problems are not conceptually as difficult for Bayesians.
- Of course, there is the problem of computing the posterior, even approximately (will MCMC be the downfall of statistics?).
- And, priors have consequences.
- So there are lots of opportunities for investigation into these consequences.

- **A Bayesian problem:** develop priors for Bayesian functional data analysis.
- Again assume the data are 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. $N(\mu, V)$:

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

- 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 $\vec{V}_{ij} = V(t_i, t_j)$.
- Prior distribution for μ: $\mu|V, k \sim N(0, kV)$.
- But $V \sim$?????
- What priors can we construct for covariance functions?

Requisite properties of covariance functions:

- Symmetry: $V(s,t) = V(t,s)$.
- Positive definiteness: for any choice of k and distinct s_1, ..., s_k in the domain, the matrix given by $\vec{V}_{ij} = V(s_i, s_j)$ is positive definite.
- It is difficult to achieve this latter requirement.

Requirements on Covariance Priors:

- Our first requirement in constructing a prior for covariance functions is that we mind the GRIP
- One may wish to use the conjugate inverse Wishart prior: $\vec{V}^{-1} \sim Wishart(d_m, W_m)$ for some $m \times m$ matrix W_m.
- ... where, e.g., W_m is obtained by discretizing a standard covariance function.
- Under what conditions (if any) on m and d_m will this converge to a probability measure on the space of covariance operators?
- This would be an indirect approach to satisfying the GRIP. More on this later.

Requirements on Covariance Priors (cont.):

- An easier 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, $\vec{V}_{ij} = V(t_i, t_j)$.
- i.e., the direct approach.
- We (joint work with Hong Xiao Zhu of MDACC) did come up with something that works, sort of.

A proposed approach that does work (sort of):

- Suppose Z_1, Z_2, ... are i.i.d. realizations of a Gaussian random process (mean 0, covariance function $B(s,t)$).
- Consider
$$V(s,t) = \sum_i w_i Z_i(s) Z_i(t)$$
where w_1, w_2, ... are nonnegative constants satisfying
$$\sum_i w_i < \infty.$$
- One can show that this gives a random covariance function, and that its distribution "fills out" the space of covariance functions.
- Can we compute with it?

A proposed approach that sort of works (cont.):

- Thus, if we can compute with this proposed prior, we will have satisfied the three requirements: a valid prior on covariance functions that "fills out the space" of covariance functions, and is useful in practice.
- 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 sort of works (cont.):

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

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

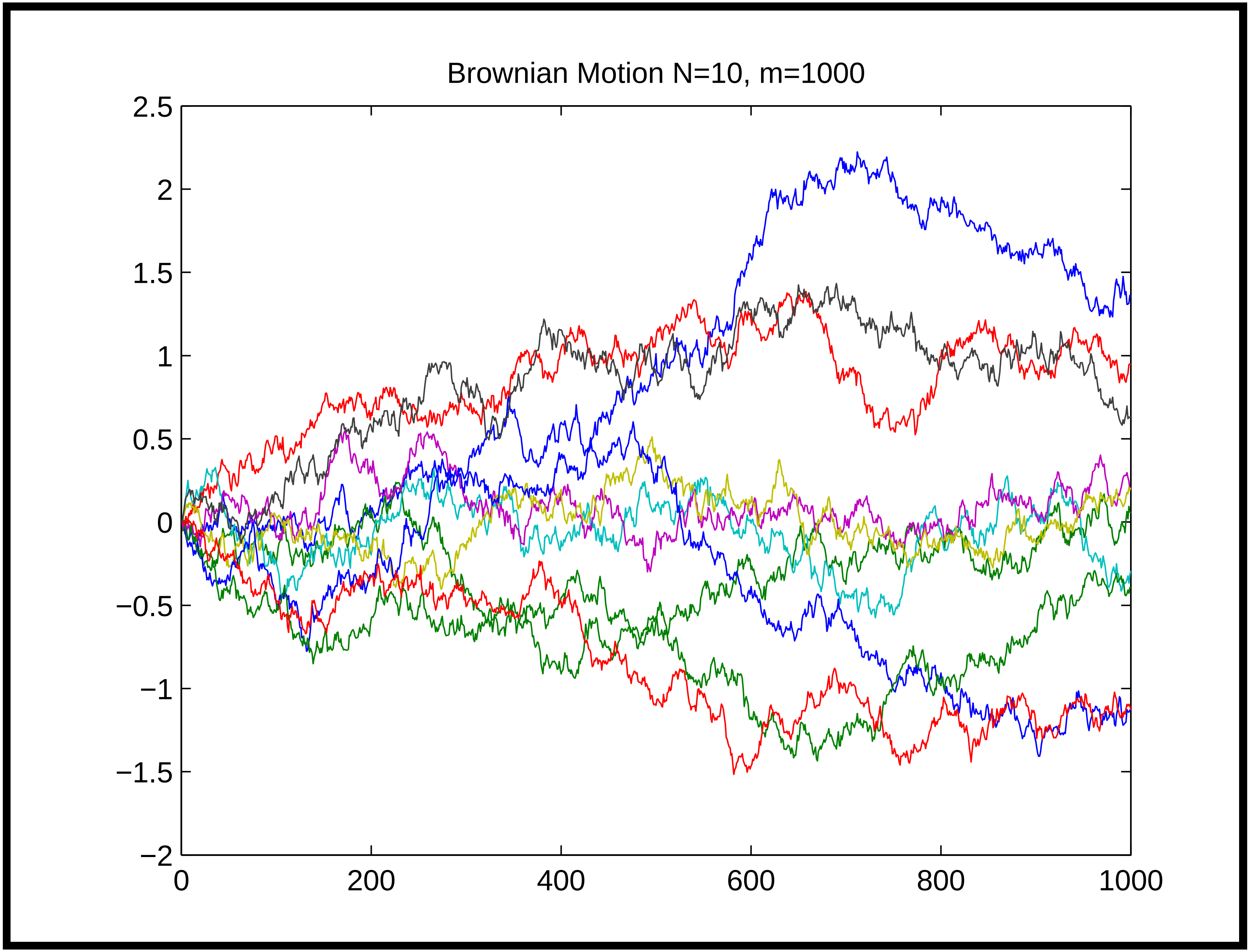
- We devised a Metropolis-Hastings algorithm to sample the Z_i.
- Can use rank-1 QR updating to do fairly efficient computing (update each $\vec{Z}_i$ one at a time).

A proposed approach that sort of works (cont.):

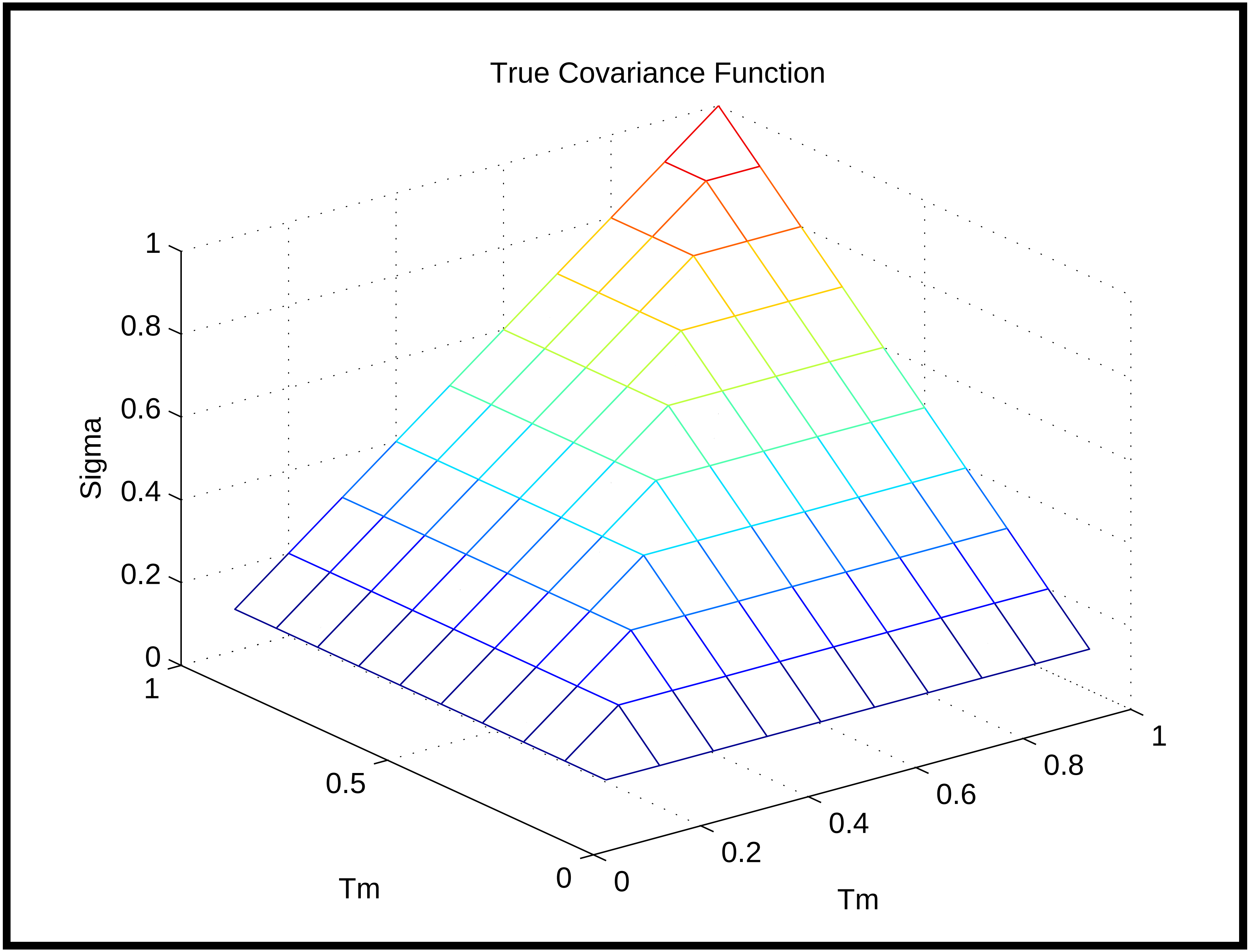
- There are a couple of minor modifications:
  1. We include an additional scale parameter k in $V(s,t) = k\sum_i w_i Z_i(s) Z_i(t)$ where k has an independent inverse Γ prior.
  2. We integrate out μ and k. and use the marginal unnormalized posterior $f(\boldsymbol{Z}|\vec{Y}_1, \ldots \vec{Y}_n)$ in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

Some results with simulated data:

- 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.

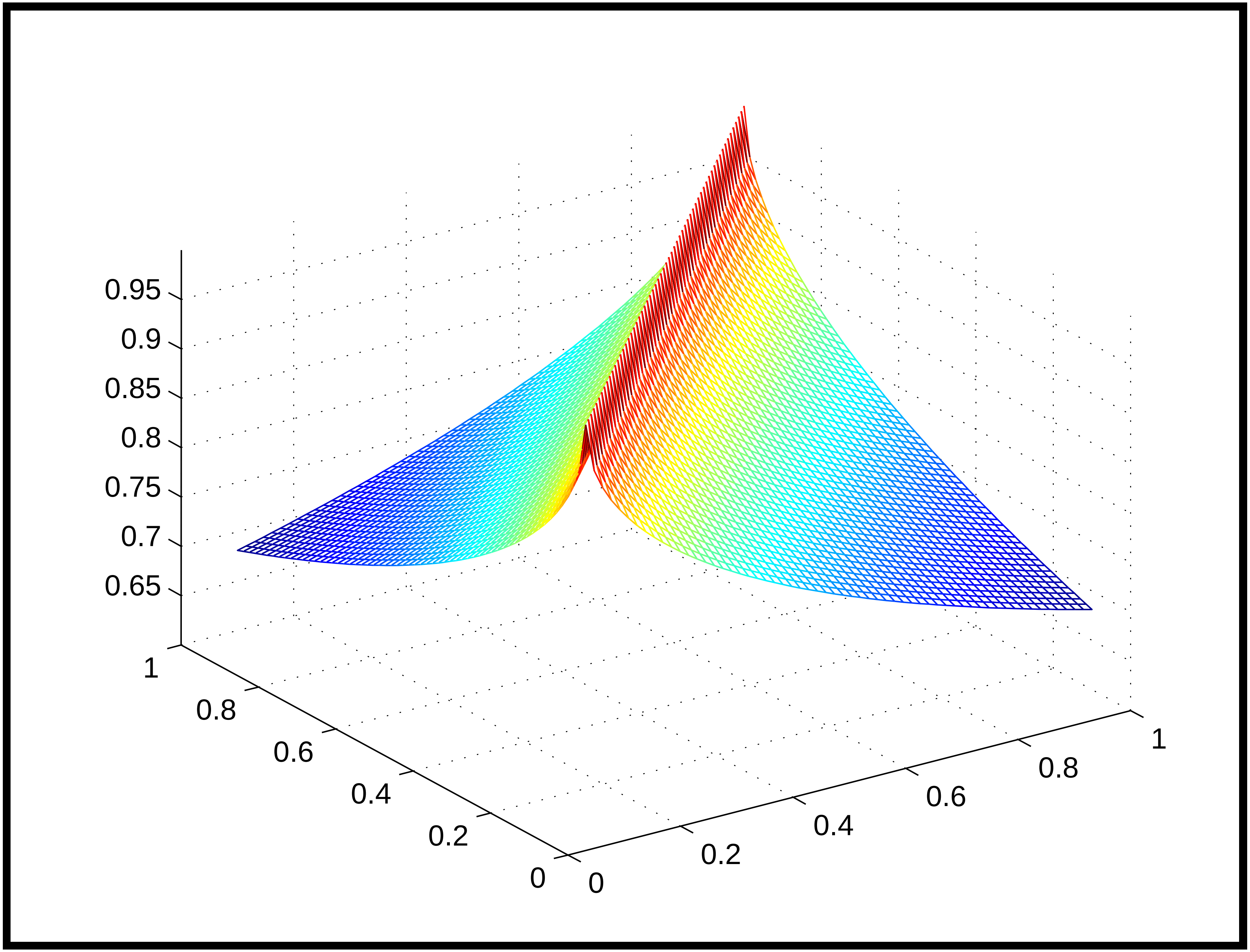


True Covariance Function
Sigma
1
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

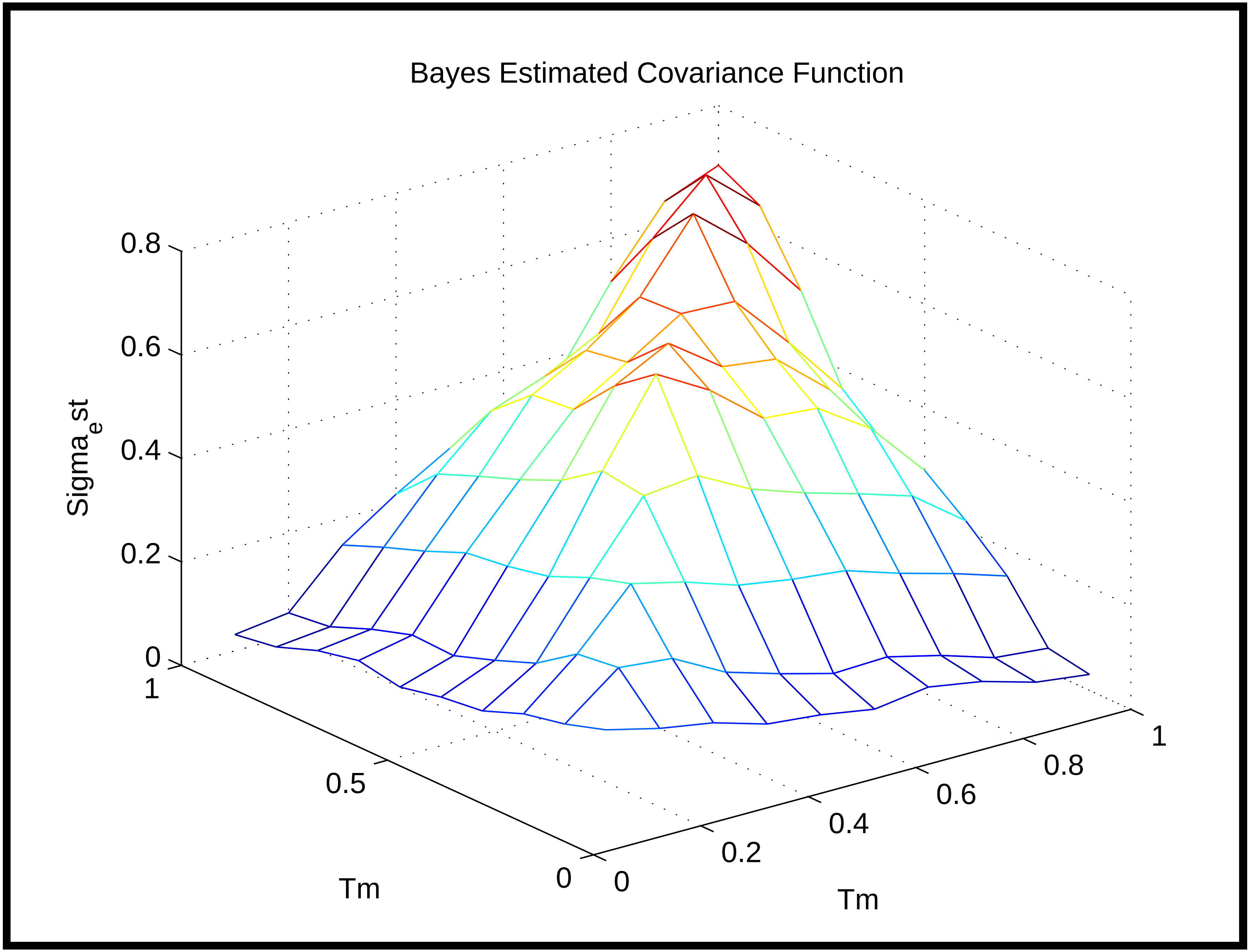
The covariance function used to generate the Z_i is the Ornstein-Uhlenbeck correlation:

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

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

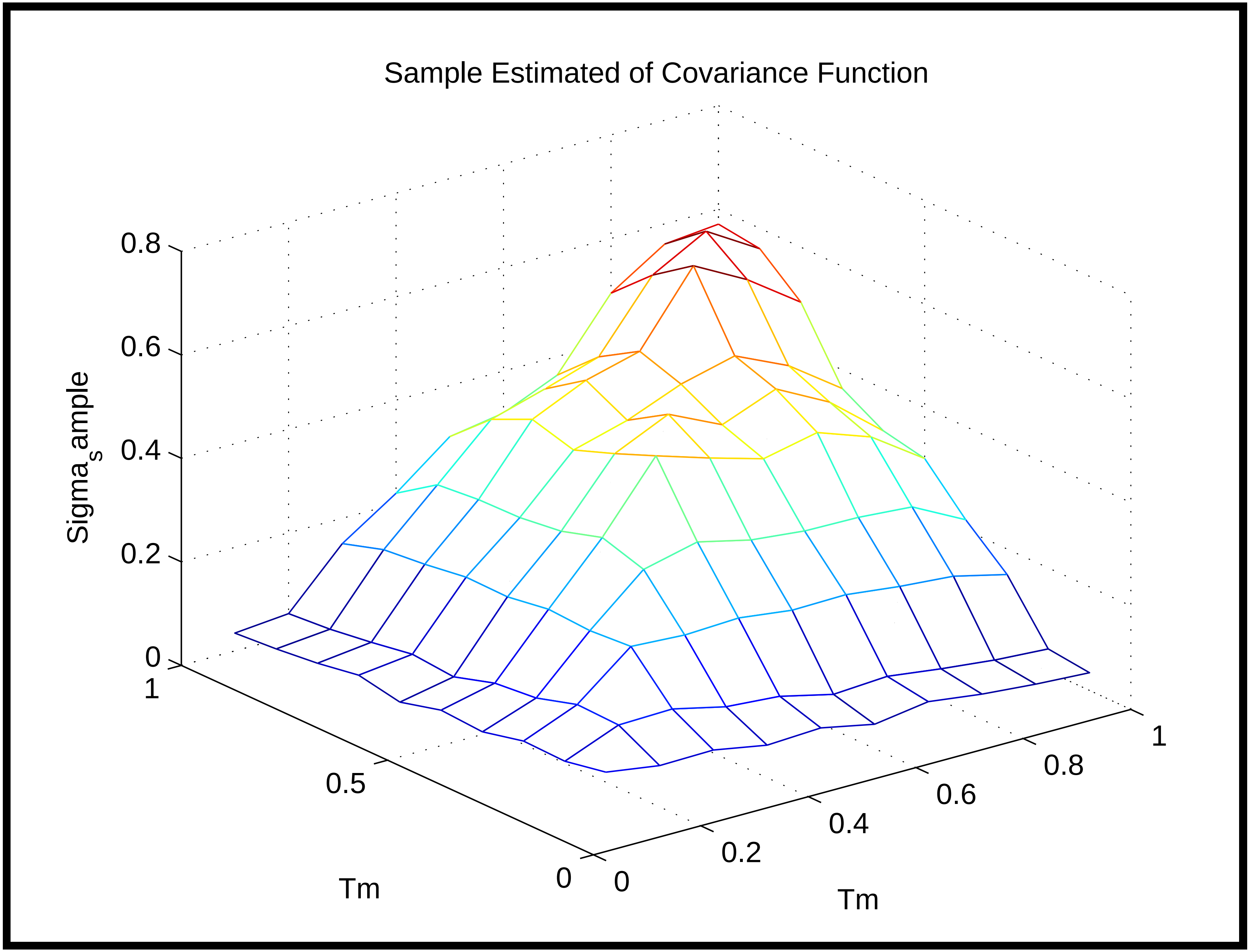

0.95
0.9
0.85
0.8
0.75
0.7
0.65
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

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



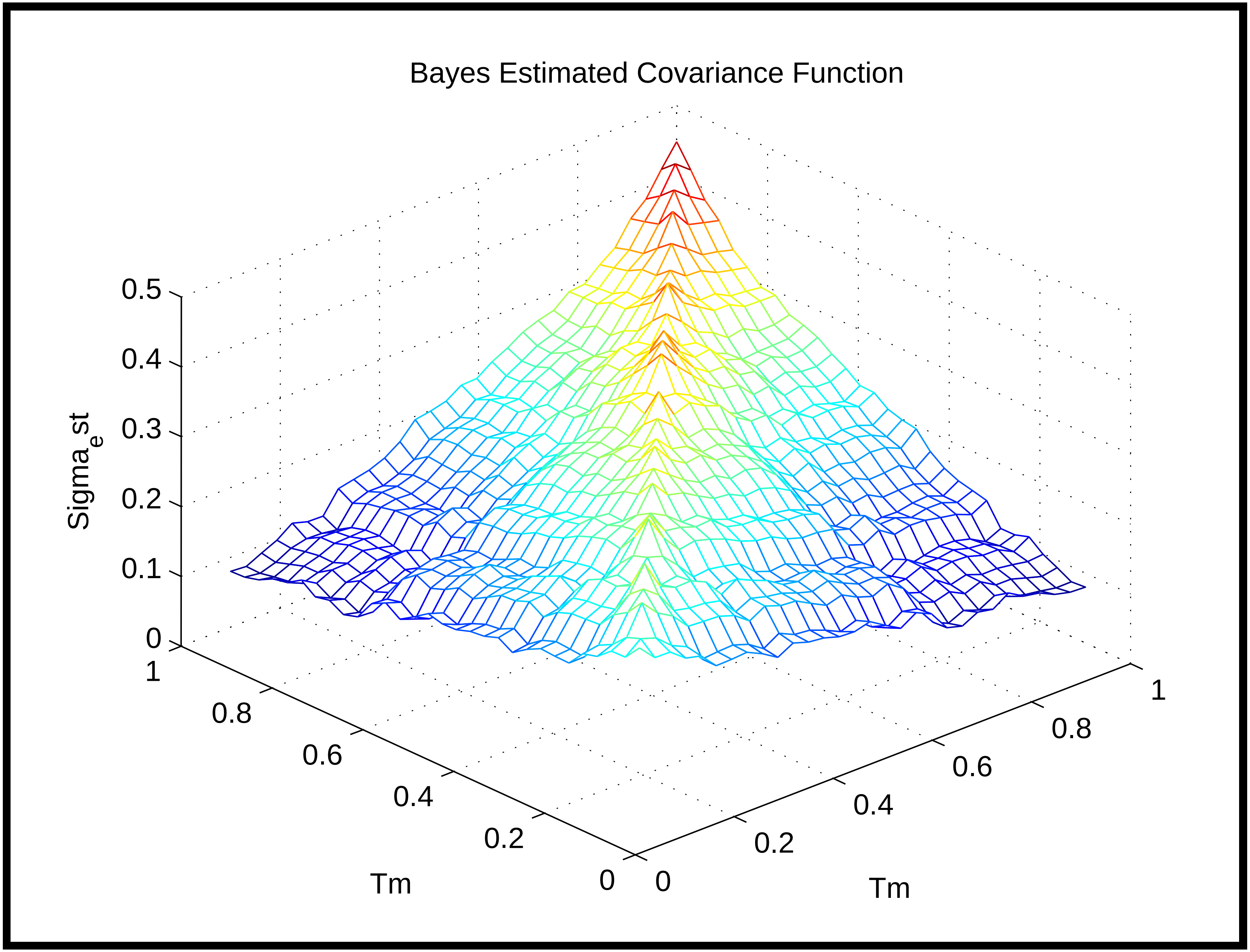
Bayes Estimated Covariance Function
Sigma_est
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

The sample covariance estimate with $m = 10$.

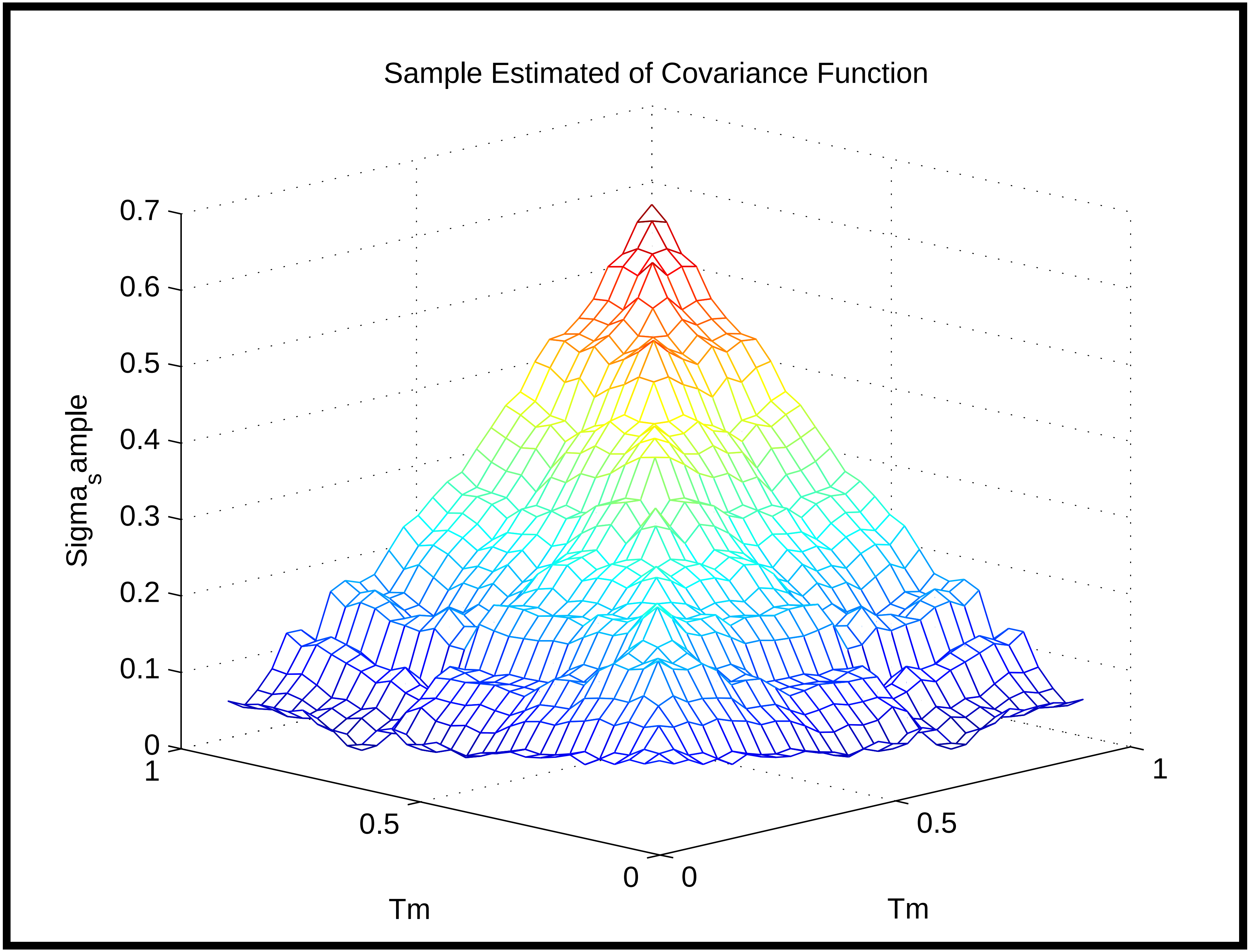


Sample Estimated of Covariance Function
Sigma_sample
0.8
0.6
0.4
0.2
0
1
0.5
0
Tm
0
0.2
0.4
0.6
0.8
1
Tm

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
0
0.5
1
Tm
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

Problems with the proposed approach that sort of works (cont.):

- The problem is way over-parameterized in terms of the $\vec{Z}_j$, $1 \leq j \leq J$, where $J \gg m$.
- Computations very time intensive, and MCMC seems to not mix well - seems to converge to different values depending on the start.
- Caused by complex non-identifiability in the model? Posterior "mode" is a complicated manifold in a very high dimensional space.

Another approach (work in progress):

- It would be very nice if we could construct a conjugate prior like the inverse Wishart in finite dimensions.
- This seems problematic. The main difficulty is that the inverse of a covariance operator (obtained from a covariance function) 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

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

- Compute $V^{-1}g$ by solving (for f) the integral equation

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

Inverse Wishart (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

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

Inverse Wishart (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 using 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 $\sum_i \lambda_i < \infty$.

Inverse Wishart (cont.):

- So, even though it looks like it is going to be very difficult to make it work, is there some way to do so?
- Instead of trying to guess a prior for which an inverse Wishart will be a good finite dimensional approximant, let's try another approach.
- Let's see if we can choose d_m so that as $m \to \infty$, $InverseWishart(d_m, \vec{B}_m)$ converges (in some sense).
- It is very difficult working with the Inverse Wishart - no m.g.f., the ch.f. is unknown.

Inverse Wishart (cont.):

- In order to obtain our results, we define "sampling" and interpolation operators:

$$
\begin{aligned}
\vec{f}_m &= (f(t_1), \ldots, f(t_m)) \\
\mathcal{I}\vec{f}_m &= \text{linear interpolant of } \vec{f}_m.
\end{aligned}
$$

- Here, $(t_1, \ldots, t_m)$ is a regular grid. Note that $f \mapsto \vec{f}_m$ is an operator from continuous functions to m-dimensional space, and $\mathcal{I}$ goes the other way.

- Define an analogous sampling operator for functions of two variables: $\vec{B}_m$ is an $m \times m$ matrix with (i, j) entry equal to $B(t_i, t_j)$.

Inverse Wishart (cont.):

- Moment results: suppose $\vec{V}_m \sim InverseWishart(d_m, s_m \vec{B}_m)$ and $\vec{f}_m$ is obtained by "sampling" a continuous function f.
- Then as long as $m/d_m \to a > 1$,

$$\frac{E[\mathcal{I}\vec{V}_m \vec{f}_m]}{d_m - m} \to Bf/(a-1),$$

where

$$Bf(s) \; = \; \int B(s,t)f(t)dt.$$

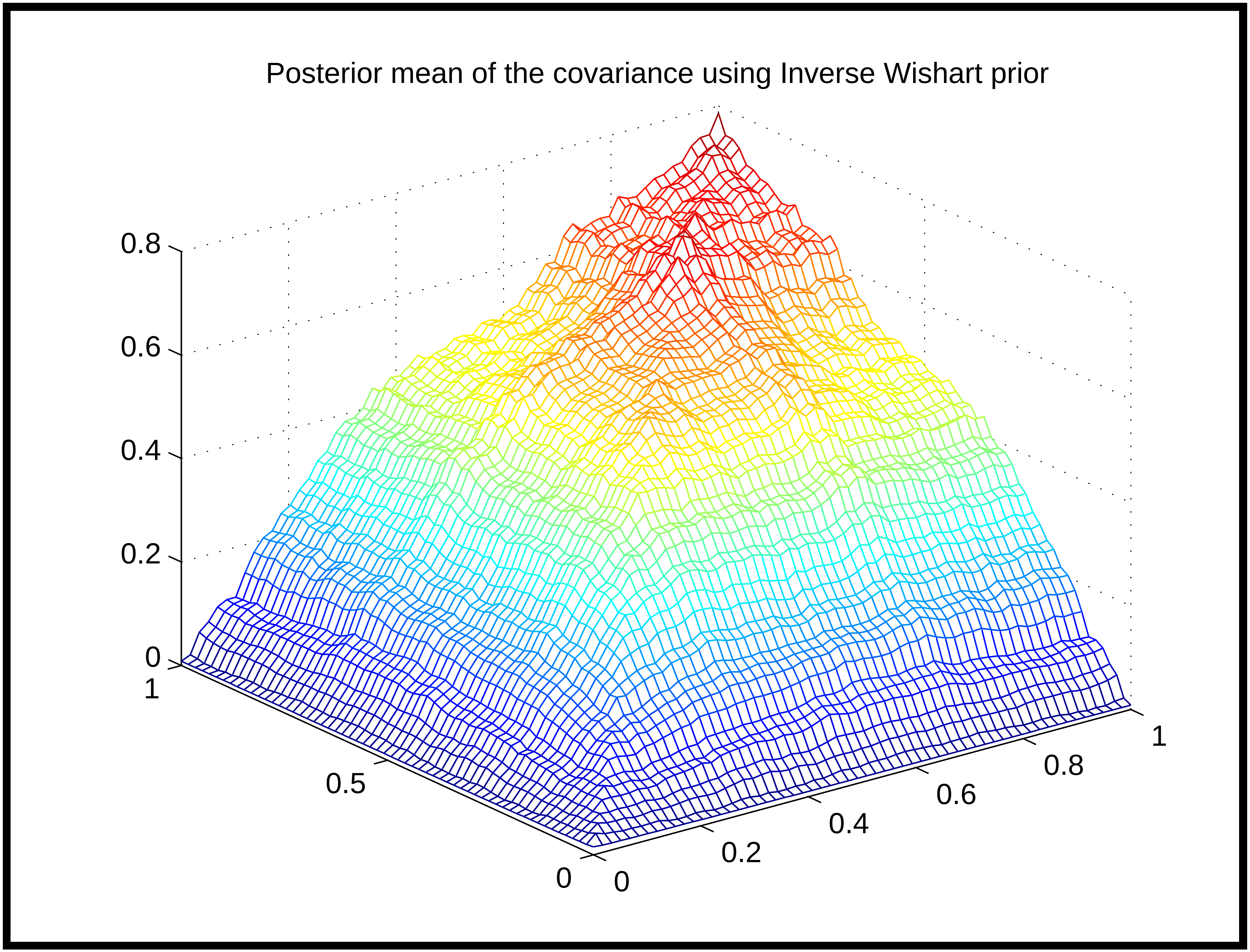
Inverse Wishart (cont.):

- Second Moments results: suppose $\vec{V}_m \sim InverseWishart(d_m, s_m\vec{B}_m)$ and $\vec{f}_m$ and $\vec{g}_m$ are obtained by "sampling" continuous functions f and g,
- Again as long as $m/d_m \to a > 1$,

$$\frac{E[\mathcal{I}\vec{V}_m \vec{f}_m \vec{g}_m^T \vec{V}_m]}{(d_m - m)^2} \to Bf \otimes Bg/(a-1)^2.$$

- Thus, in some sense, we can get first and second moments to converge if we have d_m/m converging (e.g., take $d_m = 2m$).

The Bayesian posterior mean estimate under inverse-Wishart prior with $m = 50$, $d_m = 100$ obtained by Monte-Carlo.



Posterior mean of the covariance using Inverse Wishart prior
0.8
0.6
0.4
0.2
0
1
0.5
0
0
0.2
0.4
0.6
0.8
1

Further research:

- Main interesting problem in the direct approach: find ways to approximate the prior using mixtures of inverse Wisharts.
- For the indirect approach: nearly complete proof for weak convergence in the space of S-operators but using a basis function expansion rather than grid evaluations.
- Must check the properties of this limiting measure.

The End