Stable Parameterization Schemes for Gaussians

Michael McCourt

Department of Mathematical and Statistical Sciences
University of Colorado Denver

ICOSAHOM 2014
Salt Lake City
June 24, 2014
I would like to offer my thanks to Rodrigo Platte for the invitation to this session, and to all of you in attendance for your attention.
Introduction

The shape parameter $\varepsilon$ associated with some RBFs plays an important role in the quality of associated approximations.

Example Gaussian support vector machine decision contour

- $\varepsilon = 10$, Too Erratic
- $\varepsilon = 1$, Nice
- $\varepsilon = .1$, Insensitive

Comparison with the true pattern allows for judgment.

How can the quality be judged without the true pattern?
Kernel-based Interpolation

Basic strategies for parameterizing kernels are based on:
- Error bounds ([9] through [10])
- “Stationary” interpolation strategies (e.g., [13], or [7])
- Trial and error (almost everyone)

Today, we focus on **statistical** approaches to optimizing $\varepsilon$, generally based on the interpretation of scattered data interpretation through spatial statistics or data assimilation.

We restrict our discussion to **Gaussian** kernels $K(x, z) = e^{-\varepsilon^2 \|x-z\|^2}$; most results apply to other positive definite kernels as well.
Statistical Parameterization Schemes

By using spatial statistics to analyze data, we open up avenues:

- Kriging variance, a.k.a., the power function
  - Golomb-Weinberger bound
- Maximum likelihood estimation
  - Maximum a posteriori estimation
- $k$-fold cross-validation (not discussed here)
  - Generalized cross-validation (also not discussed)

Goal

In this talk, we will look at how these quantities can be computed for small $\varepsilon$ values using a stable basis for Gaussians [3].
Hilbert-Schmidt Eigenexpansion

The term Hilbert-Schmidt SVD (HS-SVD) was introduced in [2], though the idea originated in [5].

We recall the Hilbert-Schmidt eigenvalue problem,

\[ \int K(x, z) \varphi_n(z) \rho(z) \, dz = \lambda_n \varphi_n(x), \]

which allows us to write a positive definite kernel as

\[ K(x, z) = \sum_{|n|=d} \lambda_n \varphi_n(x) \varphi_n(z), \quad n \text{ is a multiindex.} \]

By exploiting this series, the standard basis

\[ k(x)^T = (K(x, x_1) \quad \cdots \quad K(x, x_N)) \]

can be expressed using a stable basis \( \psi(x)^T \).
Hilbert-Schmidt Eigenexpansion for Gaussians

Gaussians in \( d \) dimensions have the eigenexpansion

\[
\lambda_n = \frac{\sqrt{\alpha}}{(\varepsilon^2 + \delta^2 + \alpha^2)(2|n|-d+1/2)} \varepsilon^2(|n|-d), \quad \varphi_n(x) = \gamma_n e^{\delta^2 \|x\|^2} H_{n-1}(\beta \alpha x)
\]

\(\alpha\) is a free parameter (another one ...), and \(\beta, \gamma_n, \delta^2\) are defined by \(\varepsilon\) and \(\alpha\); ignore them.

Important Limits

For small \(\varepsilon\), the eigenvalues decay quickly and the eigenfunctions approach Hermite polynomials.

\[
\lim_{\varepsilon \to 0} \lambda_n \sim \varepsilon^2(|n|-d) \quad \lim_{\varepsilon \to 0} \varphi_n(x) \sim H_{n-1}(\alpha x)
\]

Designer Kernels

Work such as [15] and [4] suggest constructing kernels by choosing eigenfunctions to satisfy certain properties.
Eigenfunction Decompositions and Matrix Manipulations

The $N \times N$ kernel matrix $K$ can be written as the infinite matrix product

$$K(x, z) = \sum_{|n|=d}^{\infty} \lambda_n \varphi_n(x) \varphi_n(z) \Rightarrow K = \Phi \Lambda \Phi^T = \Phi \begin{pmatrix} \Lambda_1 & \\
 & \Lambda_2 \end{pmatrix} \begin{pmatrix} \Phi_1^T \\
 & \Phi_2^T \end{pmatrix},$$

where $\Phi_1, \Lambda_1$ are $N \times N$ and invertible. The Hilbert-Schmidt eigenvalues appear in nonincreasing order on the diagonal of $\Lambda$.

We may write

$$K = \Phi \begin{pmatrix} \Lambda_1 \Phi_1^T \\
 & \Lambda_2 \Phi_2^T \end{pmatrix} = \Phi \begin{pmatrix} I_N \\
 & \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1} \end{pmatrix} \Lambda_1 \Phi_1^T = \Psi \Lambda_1 \Phi_1^T.$$

$K = \Psi \Lambda_1 \Phi_1^T$ is the Hilbert-Schmidt SVD, or just HS-SVD.

$\Psi$ is made of the eigenfunctions plus a “small” correction

$$\Psi = (\Phi_1 \quad \Phi_2) \begin{pmatrix} I_N \\
& \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1} \end{pmatrix} = \Phi_1 + \Phi_2 \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1}.$$
Stable Interpolation Through the Hilbert-Schmidt SVD

Forgoing the details, the standard basis and stable basis are related by

\[ k(x)^T = \psi(x)^T \Lambda_1 \Phi_1^T, \quad K = \psi \Lambda_1 \Phi_1^T. \]

\( k(x_i)^T \) and \( \psi(x_i)^T \) are the \( i \)th rows of \( K \) and \( \psi \) respectively.

Evaluating the interpolant \( s \) at the point \( x \) can now be done stably:

\[
    s(x) = k(x)^T K^{-1} y \\
    = \psi(x)^T \Lambda_1 \Phi_1^T \Phi_1^{-T} \Lambda_1^{-1} \psi^{-1} y \\
    = \psi(x)^T \psi^{-1} y.
\]

This new basis \( \psi \) is more stable than \( k \) because the \( \Lambda_1 \) term is eliminated \textit{analytically}. For small \( \varepsilon \), the condition of \( \Lambda_1 \) becomes untenable.

michael.mccourt@ucdenver.edu  (UCDenver)
More About the HS-SVD

Notes

- This strategy is often referred to as RBF-QR when $\Phi_2^T \Phi_1^{-T}$ is computed through a QR factorization.
- The vector $k(x)^T K^{-1} = \psi(x)^T \Psi^{-1}$ contains the Cardinal functions.
- We are confident in the existence of a HS-SVD. Uniqueness is another question ...

Strategy

We want to study the role that this idea plays in the objective functions associated with various parameterization schemes.

Most of our work today involves exploiting the HS-SVD to try and open up the range of viable $\varepsilon$ values for our Gaussians.
HS-SVD of $K$ is NOT a Symmetric Decomposition

Interesting Point

Despite the SPD nature of $K$, $\psi \Lambda_1 \phi_1^T$ is NOT a symmetric decomposition, i.e., $\psi \neq \phi_1$.

While surprising, this does not severely hinder analysis. To get some practice working with the HS-SVD, let us study $\psi^{-1} \phi_1$ and carefully apply inverses:

$$\psi^{-1} \phi_1 = (\Phi_1 + \Phi_2 \Lambda_2 \Phi_2^T \phi_1^{-T} \Lambda_1^{-1})^{-1} \phi_1$$

$$= (I_N + \Phi_1^{-1} \phi_2 \Lambda_2 \Phi_2^T \phi_1^{-T} \Lambda_1^{-1})^{-1}$$

$$= \Lambda_1 (\Lambda_1 + \Phi_1^{-1} \phi_2 \Lambda_2 \Phi_2^T \phi_1^{-T})^{-1}$$

$$= \Lambda_1 (\Lambda_1 + G^T G)^{-1},$$

where $G = \Lambda_2^{1/2} \phi_2 \phi_1^{-T}$. Clearly, $\psi^{-1} \phi_1$ is symmetric, since it is a product of symmetric matrices.
HS-SVD of K Approaches a Symmetric Decomposition as $\varepsilon \to 0$

We know that

$$\Psi^{-1}\Phi_1 = \Lambda_1(\Lambda_1 + G^TG)^{-1}, \quad G = \Lambda_2^{1/2}\Phi_2\Phi_1^{-T}.$$ 

We can study the magnitude of these components:

$$\lim_{\varepsilon \to 0} \|\Lambda_1\|_2 = \varepsilon^0 = 1,$$

$$\lim_{\varepsilon \to 0} \|G^TG\|_2 = \lim_{\varepsilon \to 0} \|\Lambda_2^{1/2}\Phi_2\Phi_1^{-T}\|_2^2 \approx O\left((\varepsilon N^{1/d})^2\right).$$

This suggests that $(\Lambda_1 + G^TG)^{-1} \to \Lambda^{-1}$ and

$$\Psi^{-1}\Phi_1 \to \Lambda_1\Lambda_1^{-1} = I_N, \quad \text{as } \varepsilon \to 0.$$ 

Applied to Parameterization Schemes

Enough practice ... let’s see how this really works.
Kernel-based Interpolation as Kriging

Two strategies for studying scattered data \( \{ x_i, y_i \}_{i=1}^{N} \equiv \{ \mathcal{X}, \mathcal{Y} \} \):

**Numerical Analysis**

Data was generated by a function from the RKHS induced by \( K, \mathcal{H}_K \).

\[
s(x) = k(x)^T K^{-1} y, \quad |y(x) - s(x)| \leq \|y\|_{\mathcal{H}_K} \sqrt{K(x, x) - k(x)^T K^{-1} k(x)}
\]

**Spatial Statistics**

Data was realized from a zero-mean Gaussian random field \( Y \) with covariance \( K \).

\[
Y_x | \{ \mathcal{X}, \mathcal{Y} \} \sim \mathcal{N} \left( k(x)^T K^{-1} y, K(x, x) - k(x)^T K^{-1} k(x) \right)
\]

Note that \( V(x) = P(x)^2 = K(x, x) - k(x)^T K^{-1} k(x) \).
Kriging Variance

Minimizing $V(\mathbf{x})$ would, in a sense, yield the most trustworthy predictions. Its computation can be subject to ill-conditioning as $\varepsilon \to 0$

$$V(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - \mathbf{\psi}(\mathbf{x})^T \mathbf{\psi}^{-1} \mathbf{k}(\mathbf{x})$$

Recall the Cardinal functions: $\mathbf{k}(\mathbf{x})^T K^{-1} = \mathbf{\psi}(\mathbf{x})^T \mathbf{\psi}^{-1}$.

Better stability, but numerical cancelation is unavoidable.
Golomb-Weinberger Bound

The standard native space error bound can be augmented into the Golomb-Weinberger bound [6]

$$|y(x) - s(x)| \leq \|y\|_{\mathcal{H}_K} \sqrt{V(x)} \leq \delta \|s\|_{\mathcal{H}_K} \sqrt{V(x)},$$

for some small $\delta$ value related to the quality of the interpolation.

Ignoring the fact that $V(x)$ is still limited by cancelation, we must also compute $\|s\|_{\mathcal{H}_K}$ successfully to evaluate this bound. Recall that, through the reproducing property,

$$\|s\|_{\mathcal{H}_K}^2 = \langle s, s \rangle_{\mathcal{H}_K} = \langle y^T K^{-1} k(\cdot), k(\cdot)^T K^{-1} y \rangle_{\mathcal{H}_K}$$

$$= y^T K^{-1} \langle k(\cdot), k(\cdot)^T \rangle_{\mathcal{H}_K} K^{-1} y$$

$$= y^T K^{-1} KK^{-1} y = y^T K^{-1} y.$$
The Hilbert Space Norm (Mahalanobis Distance) $y^T K^{-1} y$

Define $\Psi b = y$, the system for the stable basis interpolation coefficients. Using this and $K = \Psi \Lambda_1 \Phi^T$ in $y^T K^{-1} y$ gives

$$y^T K^{-1} y = b^T \Psi^T \Phi_1^{-T} \Lambda_1^{-1} \Psi^{-1} \Psi b = b^T (\Psi^{-1} \Phi_1)^{-T} \Lambda_1^{-1} b$$

Our old friend $\Psi^{-1} \Phi_1 = (I_N + \Phi_1^{-1} \Phi_2 \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1})^{-1}$ returns! Canceling out the double inverse and applying the transpose gives

$$y^T K^{-1} y = b^T (I_N + \Lambda_1^{-1} \Phi_1^{-1} \Phi_2 \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1}) \Lambda_1^{-1} b$$

$$= b^T \Lambda_1^{-1} b + b^T \Lambda_1^{-1} \Phi_1^{-1} \Phi_2 \Lambda_2 \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1} b$$

$$= b^T \Lambda_1^{-1} b + (\Lambda_2^{1/2} \Phi_2 \Phi_1^{-T} \Lambda_1^{-1} b)^T (\Lambda_2^{1/2} \Phi_2 \Phi_1^{-T} \Lambda_1^{-1} b)$$

$$\geq b^T \Lambda_1^{-1} b$$

Computing this quantity is feasible, although it grows quickly. The bound in the last line “may be” rather tight for small $\varepsilon$. 

michael.mccourt@ucdenver.edu (UCDenver)
The Hilbert Space Norm - A “Tight” Bound

Again considering Gaussians with small $\varepsilon$,

$$
\lim_{\varepsilon \to 0} \left\| \Lambda_2^{1/2} \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1} b \right\|_2^2 \leq \lim_{\varepsilon \to 0} \left\| \Lambda_2^{1/2} \Phi_2^T \Phi_1^{-T} \right\|_\infty \left\| \Lambda_1^{-1} \right\|_\infty \left\| b \right\|_2
$$

This lets us say that

$$
\lim_{\varepsilon \to 0} \left\| \Lambda_2^{1/2} \Phi_2^T \Phi_1^{-T} \Lambda_1^{-1} b \right\|_2^2 \sim \varepsilon^{2-2N^1/d} \left\| b \right\|_2^2.
$$

as compared to

$$
b^T \Lambda_1^{-1} b = \left\| \Lambda_1^{-1/2} b \right\|_2^2 \leq \left\| \Lambda_1^{-1/2} \right\|_\infty^2 \left\| b \right\|_2^2 = \varepsilon^{-2N^1/d} \left\| b \right\|_2^2.
$$

Thus we may compute, or at least bound, $\left\| s \right\|_{\mathcal{H}_K} = y^T K^{-1} y$. 
Maximum Likelihood Estimation

Even though cancelation in the Kriging variance prevents us from leveraging $\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}$ in the Golomb-Weinberger bound, it can be used in the maximum likelihood estimator [11].

Saving some time, we write that the MLE $\varepsilon$ is the $\varepsilon$ which minimizes

$$C_{\text{MLE}}(\varepsilon) = N \log(\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}) + \log \det \mathbf{K}. $$

Using the HS-SVD, we know that

$$\log \det \mathbf{K} = \log \det(\Psi \Lambda_1 \Phi_1^T) = \log \det \Psi + \log \det \Lambda_1 + \log \det \Phi_1. $$

This gives us a tool for parameterizing Gaussians with small $\varepsilon$. 
MLE Example

Example: \( N = 169 \) Halton points fit to
\[
\log(3 + x + y) + 2(1 + x + y)^2 + 2 \cos(2xy) \text{ on } [-1, 1]^2.
\]

Primary bound: \( b^T \Lambda^{-1} b \)

Correction: \( \left\| \Lambda_2^{1/2} \Phi_2 \Phi_1^{-T} \Lambda_1^{-1} b \right\|_2^2 \)
Maximum A Posteriori Estimation

When prior beliefs regarding an “optimal” $\varepsilon$ value are available, they can supplement the likelihood function.

This idea is discussed in general in, e.g., [8], and in the context of Gaussian Processes in [12] and [14].
Conclusion

Results

Repeating the strategy for interpolation, the HS-SVD may provide opportunities to study Gaussian parameterization schemes for small $\varepsilon$.

- The Hilbert space norm, MLE and Cross-validation (not discussed) are viable for small $\varepsilon$.
- Other PD kernels will work as well if their eigenexpansion is available or can be approximated.

To Be Accomplished

- Kriging variance is better but still troubled.
- Reduce computational cost.
- Consider incorporation with matrix-free approaches, e.g., [1].
- Tests on applications.
Appendix: $k$-fold Cross-Validation

Cross-Validation Strategy

Using subsets of your data $\mathcal{X}_I$, predict values at omitted points $\mathcal{X}_O$ and use the residual to measure the quality of that parameterization.

If we block up our kernel system $K \mathbf{c} = \mathbf{y}$ matrix as

$$K = \begin{pmatrix} K_{II} & K_{IO} \\ K_{OI} & K_{OO} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_I \\ y_O \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_I \\ c_O \end{pmatrix},$$

such that $K \mathbf{a} = \mathbf{l}$, the residual at $\mathcal{X}_O$ is $\| y_O - K_{OI} K_{II}^{-1} y_I \|$. Summing all the residuals over $k$ disjoint omitted point sets $\mathcal{O} = \{ \mathcal{X}_O^{(1)}, \ldots, \mathcal{X}_O^{(k)} \}$ would produce

$$C_{CV}(\varepsilon; \mathcal{O}) = \sum_{\mathcal{X}_O \in \mathcal{O}} \| y_O - K_{OI} K_{II}^{-1} y_I \|.$$

This is what we want to minimize, but $K_{II}^{-1}$ may be dangerous.
Appendix: $k$-fold Cross-Validation with HS-SVD

Rewriting the blocks of $K$ with the HS-SVD gives

$$K = \begin{pmatrix} K_{II} & K_{IO} \\ K_{OI} & K_{OO} \end{pmatrix} = \begin{pmatrix} \Psi_{II} \Lambda_{I} \Phi_{I}^{T} & \Psi_{IO} \Lambda_{O} \Phi_{O}^{T} \\ \Psi_{OI} \Lambda_{I} \Phi_{I}^{T} & \Psi_{OO} \Lambda_{O} \Phi_{O}^{T} \end{pmatrix}$$

Notably, it says $K_{II} = \Psi_{II} \Lambda_{I} \Phi_{I}^{T}$ and $K_{OI} = \Psi_{OI} \Lambda_{I} \Phi_{I}^{T}$. Substituting into the residual $\left| y_{O} - K_{OI} K_{II}^{-1} y_{I} \right|$

$$K_{OI} K_{II}^{-1} = \Psi_{II} \Lambda_{I} \Phi_{I}^{T} \Phi_{I}^{-T} \Lambda_{I}^{-1} \Psi_{II}^{-1} = \Psi_{OI} \Psi_{II}^{-1}$$

gives the stable residual (unsurprisingly) as $\left| y_{O} - \Psi_{OI} \Psi_{II}^{-1} y_{I} \right|$. 
Appendix: Cross-Validation, Rippa Trick

Define the blocked kernel system, $K\mathbf{c} = y \Rightarrow \mathbf{c} = Ay$:

$$K = \begin{pmatrix} K_{II} & K_{IO} \\ K_{OI} & K_{OO} \end{pmatrix}, \quad A = \begin{pmatrix} A_{II} & A_{IO} \\ A_{OI} & A_{OO} \end{pmatrix}, \quad y = \begin{pmatrix} y_I \\ y_O \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_I \\ c_O \end{pmatrix}.$$  

This says that $c_O = A_{OI}y_I + A_{OO}y_O$, and $AK = I$ gives us $A_{OI} = -A_{OO}K_{IO}^{-1}K_{II}$, so

$$c_O = -A_{OO}K_{OI}K_{II}^{-1}y_I + A_{OO}y_O,$$

$$A_{OO}^{-1}c_O = -K_{OI}K_{II}^{-1}y_I + y_O.$$  

Thus, the residual $|y_O - K_{OI}K_{II}^{-1}y_I| = |A_{OO}^{-1}c_O|.$
Appendix: Cross-Validation, Rippa Trick for HS-SVD

Rewriting the blocks of $K$ with the HS-SVD gives

$$K = \begin{pmatrix} K_{II} & K_{IO} \\ K_{OI} & K_{OO} \end{pmatrix} = \begin{pmatrix} \Psi_{II} & \Psi_{IO} \\ \Psi_{OI} & \Psi_{OO} \end{pmatrix} \begin{pmatrix} \Lambda_I \Phi_I^T \\ \Lambda_O \Phi_O^T \end{pmatrix},$$

so, by absorbing the $\Lambda \Phi$ matrix into $c$, $Kc = y$ becomes

$$\begin{pmatrix} \Psi_{II} & \Psi_{IO} \\ \Psi_{OI} & \Psi_{OO} \end{pmatrix} \begin{pmatrix} b_I \\ b_O \end{pmatrix} = \begin{pmatrix} y_I \\ y_O \end{pmatrix} \Rightarrow \begin{pmatrix} b_I \\ b_O \end{pmatrix} = \begin{pmatrix} A_{II} & A_{IO} \\ A_{OI} & A_{OO} \end{pmatrix} \begin{pmatrix} y_I \\ y_O \end{pmatrix}$$

where $B\Psi = I$. The absorption gives us $c_O = \Phi_O^{-T}\Lambda_O^{-1}b_O$ and, if you stare long enough, you can show $A_{OO} = \Phi_O^{-T}\Lambda_O^{-1}B_{OO}$, so $|A_{OO}^{-1}c_O| = |B_{OO}^{-1}b_O|$.
Appendix: Kriging Variance - A Deeper Exploration

The HS-SVD helps avoid instability, but cannot circumvent numerical cancelation - thus the limit towards machine precision.

This brief example shows the basic idea behind using the HS-SVD; now another example to more efficiently compute $V(x)$.

Using $k(x)^T = \psi(x)^T \Lambda_1 \Phi_1^T$ and $K = \psi \Lambda_1 \Phi_1^T$ in $k(x)^T K^{-1} k(x)$ gives

$$k(x)^T K^{-1} k(x) = \left( \psi(x)^T \Lambda_1 \Phi_1^T \right) \left( \psi \Lambda_1 \Phi_1^T \right)^{-1} \Phi_1 \Lambda_1 \psi(x)$$

$$= \psi(x)^T \psi^{-1} \Phi_1 \Lambda_1 \psi(x)$$

Aha! We recognize $\psi^{-1} \Phi_1$ from our earlier practice (practice makes perfect).
Appendix: Kriging Variance - Fun With Matrices
We use the result, $\Psi^{-1}\Phi_1 = \Lambda_1 (\Lambda_1 + \Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T\Phi_1^{-T})^{-1}$, in our current expression:

$$k(x)^TK^{-1}k(x) = \psi(x)^T\Psi^{-1}\Phi_1\Lambda_1\psi(x)$$

$$= \psi(x)^T\Lambda_1 (\Lambda_1 + \Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T\Phi_1^{-T})^{-1}\Lambda_1\psi(x)$$

We can pull $\Lambda_1^{1/2}$ from both sides of the inverse to write this as:

$$k(x)^TK^{-1}k(x) = \psi(x)^T\Lambda_1^{1/2}(I_N + \Lambda_1^{-1/2}\Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1/2})^{-1}\Lambda_1^{1/2}\psi(x)$$

$$= \psi(x)^T\Lambda_1^{1/2}(I_N + LL^T)^{-1}\Lambda_1^{1/2}\psi(x)$$

where $L = \Lambda_2^{1/2}\Phi_2\Phi_1^{-1}\Lambda_1^{-1/2}$. As $\epsilon \to 0$, for Gaussians,

$$\|LL^T\|_2 \leq \|L\|_2^2 \leq \|\Lambda_2^{1/2}\Phi_2\Phi_1^{-1}\|_2^2\|\Lambda_1^{-1/2}\|_2^2 \sim \epsilon^{2N^{1/d}}\epsilon^{-2(N^{1/d}d-1)} = \epsilon^2.$$ 

So for small $\epsilon$, a von Neumann series allows us to write

$$k(x)^TK^{-1}k(x) \approx \psi(x)^T\Lambda\psi(x) - \psi(x)^T\Lambda^{1/2}LL^T\Lambda^{1/2}\psi(x) + ...$$
Appendix: Kriging Variance - Reduced Computation

So, for small $\varepsilon$, we can write

$$k(x)^T K^{-1} k(x) \approx \psi(x)^T \Lambda \psi(x) - \psi(x)^T \Lambda^{1/2} LL^T \Lambda^{1/2} \psi(x) + ...$$

This allows us to evaluate the Kriging variance for small $\varepsilon$ with

$$V(x) = K(x, x) - k(x)^T K^{-1} k(x)$$

$$\approx K(x, x) - \psi(x)^T \Lambda_1 \psi(x) + \psi(x)^T \Lambda^{1/2} LL^T \Lambda^{1/2} \psi(x) - ...$$

After forming the stable basis $\psi$, no further matrix inverses (such as $\psi(x)^T \psi^{-1}$) need be computed. Of course, this will not bypass the numerical cancelation, but it can reduce the cost.


B. Fornberg and C. Piret. 
A stable algorithm for flat radial basis functions on a sphere. 

M. Golomb and H. F. Weinberger. 
Optimal approximation and error bounds. 

A. Iske. 

J. S. Liu. 
*Monte Carlo strategies in scientific computing*. 

L.-T. Luh.
The shape parameter in the Gaussian function.


W. R. Madych.
Miscellaneous error bounds for multiquadric and related interpolators.


Maximum likelihood estimation of models for residual covariance in spatial regression.


B. Nagy, J. L. Loeppky, and W. J. Welch.
Fast Bayesian inference for Gaussian process models.

