On the condition number anomaly of Gaussian correlation matrices

Zimmermann, Ralf

Published in:
Linear Algebra and Its Applications

Publication date:
2015

Document version
Peer reviewed version

Citation for published version (APA):
On the condition number anomaly of Gaussian correlation matrices

R. Zimmermann

Institute of Computational Mathematics, TU Braunschweig, 38100, Germany

Abstract

Spatial correlation matrices appear in a large variety of applications. For example, they are an essential component of spatial Gaussian processes, also known as spatial linear models or Kriging estimators, which are powerful and well-established tools for a multitude of engineering applications such as the design and analysis of computer experiments, geostatistical problems and meteorological tasks. In radial basis function interpolation, Gaussian correlation matrices arise frequently as interpolation matrices from the Gaussian radial kernel function. In the field of data assimilation in numerical weather prediction, such matrices arise as background error covariances. Over the past thirty years, it was observed by several authors from several fields that the Gaussian correlation model is exceptionally prone to suffer from ill-conditioning, but a quantitative theoretical explanation for this anomaly was lacking. In this paper, a proof for the special position of the Gaussian correlation matrix is given. The theoretical findings are illustrated by numerical experiment.

© 2014 Published by Elsevier Inc.

Keywords: Gaussian process, Condition number, Design and analysis of computer experiments, Euclidean distance matrix, Correlation matrix, Kriging, Radial basis functions, Data assimilation, Spatial linear model

Article history:
Received 25 October 2013
Accepted 22 October 2014
Submitted by Z. Strakoš

MSC: 15B99

E-mail address: ralf.zimmermann@tu-bs.de.

http://dx.doi.org/10.1016/j.laa.2014.10.038
0024-3795/© 2014 Published by Elsevier Inc.
1. Introduction

The spatial linear model in the classical setting (see [1–3]) is defined as follows. Consider a real-valued covariance stationary Gaussian process in \( d \in \mathbb{N} \) spatial dimensions

\[
y(x) = f(x)^T \beta + \epsilon(x); \quad \epsilon(x) \sim N(0, \sigma^2), \quad x \in \mathbb{R}^d,
\]

where \( f(x) = (f_1(x), \ldots, f_p(x))^T \) is the regressor vector and \( \beta = (\beta_1, \ldots, \beta_p) \) is the vector of regression coefficients. Let \( \{x^1, \ldots, x^n\} \subset \mathbb{R}^d \) be a set of mutually distinct sample points. Suppose that the stationary covariance structure is modeled via a positive definite covariance function \( \text{cov}(y(x^i), y(x^j)) = \sigma^2 \rho(\theta, (x^i - x^j)) \), by convention parametrized such that for \( v \in \mathbb{R}^d \setminus \{0\} \),

\[
\rho(\theta, v) \rightarrow \begin{cases} 
1, & \text{for } \|\theta\| \to 0 \\
0, & \text{for } \|\theta\| \to \infty,
\end{cases}
\]

with coordinate-wise range-parameters \( \theta = (\theta_1, \ldots, \theta_d) \), also referred to as the model’s hyper-parameters. The reciprocal values \( 1/\theta_k \) are called the correlation lengths. For a vector \( Y = (y(x^1), \ldots, y(x^n))^T \in \mathbb{R}^n \) of \( n \) observations, let \( R := (\rho(\theta, (x^i - x^j)))_{ij} \in \mathbb{R}^{n \times n} \) be the corresponding correlation matrix and \( r(x) := (\rho(\theta, (x^i - x^j)))_{i} \in \mathbb{R}^n \). The best linear unbiased predictor is

\[
\hat{y}(x) = f(x)^T \beta + r(x)^T R^{-1} (Y - F \beta),
\]

where the matrix \( F \) features the regressor vectors \( f(x_i)^T \) \( (i = 1, \ldots, n) \) as rows and \( \beta = (F^T R^{-1} F)^{-1} F^T R^{-1} Y \) is the generalized least-squares solution to the regression problem. Up to an additive constant, the profile log likelihood is

\[
L(Y, \theta) = -\frac{1}{2} n \log(\sigma^2(\theta)) + \log(\det(R(\theta))),
\]

(1)

where \( \sigma^2(\theta) = 1/n(Y - F \beta)^T R^{-1}(\theta)(Y - F \beta) \). Both the predictor and the likelihood function require to compute the inverse of the correlation matrix, which may be replaced by solving linear systems of the form \( RV = b \).

Closely related to Kriging is radial basis function (RBF) interpolation [4,5]. Here, Gaussian correlation matrices arise frequently as interpolation matrices, also referred to as distance matrices, and, as in Kriging, it is required to solve linear systems featuring such matrices as operators.

In the field of data assimilation for numerical weather prediction [6,7], the inverses of spatial correlation matrices appear in the area-defining optimization problem, see e.g. [7, Eq. (1)].

As a consequence, the accuracy and numerical robustness in all of the aforementioned applications depend crucially on the correlation matrices’ condition numbers, a fact that has been acknowledged by several authors and is still subject to ongoing investigations.

Please cite this article in press as: R. Zimmermann, On the condition number anomaly of Gaussian correlation matrices, Linear Algebra Appl. (2014), http://dx.doi.org/10.1016/j.laa.2014.10.038
In this regard, we relate to the following, non-exhaustive selection of papers: Diamond and Armstrong [8] prove error estimates under perturbation of covariance models, demonstrating a strong dependence on the correlation matrix’ condition number. Posa [9] investigates numerically the behavior of the associated condition number for different covariance models and varying hyper-parameters. In the same setting, Ababou et al. [10] show that likelihood-optimized hyper-parameters may correspond to ill-conditioned correlation matrices. An extensive experimental study of the condition number as a function of all parameters in the Kriging exercise is provided in [11]. The connection between the condition number and the maximum likelihood function depending on the range parameters is investigated in [12].

The effect of ill-conditioning may be countered by numerical regularization, i.e. replacing $R$ by $R + \delta I$ for a small number $\delta > 0$. In the geostatistical community, this is usually referred to as introducing a nugget in the model, see [13]. Nugget selection and its effects on the predictor and on the maximum likelihood estimation process is investigated in [14] and [15]. Estimating the covariance matrix subject to a constraint on the condition number is considered in [16].

It has been observed in several applications that the Gaussian correlation model is particularly prone to suffer from ill-conditioning, see e.g. [8, Example 2], [9, Fig. 2], [17]. The Gaussian model is the special case of $p = 2$ for the family of exponential correlation functions

$$
\rho_{\exp,p}(\theta, v) = \exp \left( - \sum_{k=1}^{d} \theta_k |v_k|^p \right), \quad 1 \leq p \leq 2.
$$

Baxter briefly discusses some of the negative properties of Gaussian correlation matrices in the context of RBF interpolation in [5, p. 9]. In [18] the special structure of the Gaussian correlation matrix was investigated for equidistant multivariate observations.

The exceptionally high conditioning of the Gaussian correlation model was termed the Gaussian anomaly in [17] and has been explored numerically in [19] in the setting of a univariate process based on uniform sample data. In this case, explicit expressions for the eigenvalues and eigenvectors of the circulant Gaussian correlation matrices exist, see [6, p. 5, eqs. (10)–(14)]. However, these formula seem to be ill-suited to express the condition number growth rate in terms of the correlation length scale. The references [4, §4.3] and [5] feature condition number formulae and estimates for equally spaced grids, in which case the radial distance matrices are Toeplitz matrices. In [4, Thm. 5.12], condition number estimates for multiquadric and polyharmonic spline radial distance matrices for scattered sample data depending on the number of sample points and the maximum distance between distinct sample locations are given. Note that in the RBF context and in data assimilation, the correlation matrices usually depend only on a single univariate length-scale parameter, even when considering scattered multivariate sample points.
In the work at hand, a linear algebra proof is presented that explains why the Gaussian anomaly must occur even in the multivariate scattered setting and, just as important, why it does generally not occur for the other members of the family of exponential correlation functions. It turns out that the key difference between the Gaussian model and the other members of the exponential correlation family in quantifying the condition number growth is that the derivative of the Gaussian correlation matrix with respect to the range parameter is rank deficient, being a standard Euclidean distance matrix \([20,21]\), while the derivatives of the exponential correlation matrices are regular for generic sample point sets. Thus, all phenomena observed in \([19]\) are explained theoretically.

The proof makes use of the important feature, that the exponential correlation family \((2)\) depends real analytically on the range-parameters \(\theta_k\), regardless the choice of the exponent \(p \in [1,2]\).

# 2. Preparatory remarks on the family of exponential correlation matrices

The family of exponential correlation functions \((2)\) was subject to ongoing investigations ever since the pioneering work of Schoenberg \([22]\) (who had a different motivation for his investigations).

This section starts with a collection of known results, some of which will be used as tools in the following. Let \(x^1, \ldots, x^n \in \mathbb{R}^d\) be a set of arbitrary mutually distinct sample points and let \(R_p(\theta) := (\rho_{\exp,p}(\theta, x^i-x^j))_{i,j \leq n} \in \mathbb{R}^{n \times n}\) be the corresponding correlation matrix according to \((2)\), \(1 \leq p \leq 2\).

## 2.1. On the positive definiteness of the family of exponential correlation matrices

For \(\theta \in \mathbb{R}^d_{>0}\), the matrix \(R_p(\theta)\) is positive definite. This follows from Schoenberg’s results \([22, \text{Corollary 3} \& \S 6]\). Note that for fixed \(\theta \in \mathbb{R}^d_{>0},\) it holds that \(\sum_{k=1}^d \theta_k |x^i_k - x^j_k|^p = \|\tilde{x}^i - \tilde{x}^j\|_p^p\) for the scaled data set \(\tilde{x}^i = ((\theta_1)\frac{1}{p}x^1_i, \ldots, (\theta_d)\frac{1}{p}x^d_i) (i = 1, \ldots, n)\). Therefore, Schoenberg’s result applies directly to \(R_p(\theta)_{i,j} = (\exp(-\|\tilde{x}^i - \tilde{x}^j\|_p))_{i,j}\).

For a more comprehensive (and more modern) account on positive correlation functions and additional references, the reader is referred to \([5, \S 2.2, \S 2.3], [4, \text{Proposition 2.1 et seq}., [23, \S 4.1, \S 4.2], [24, \text{Example 11]}\).

## 2.2. On the rank of the derivatives of the exponential correlation matrices in \(\theta = 0\)

Let \(\theta : \mathbb{R} \ni \tau \mapsto \theta(\tau) \in \mathbb{R}^d\) be a real analytic curve in \(\mathbb{R}^d\) such that \(\theta(\tau) \in \mathbb{R}^d_{>0}\) for \(\tau > 0\) passing with non-zero velocity through \(\theta(0) = 0 \in \mathbb{R}^d\), i.e. \(\theta'(0) = h \in \mathbb{R}^d_{>0}\). (The simplest such curve is \(\tau \mapsto \tau h\).) The derivative of the correlation matrix by \(\tau\) is

\[
\frac{d}{d\tau} R_p(\theta(0)) = -\left( \sum_{k=1}^d h_k |x^i_k - x^j_k|^p \right)_{i,j \leq n} = -\left( \|\tilde{x}^i - \tilde{x}^j\|_p^p \right)_{i,j \leq n} \in \mathbb{R}^{n \times n},
\]
where $\tilde{x}^i = (h_1^{1/p}x_1^i, \ldots, h_d^{1/p}x_d^i)^T$ ($i = 1, \ldots, n$). Hence, the derivative of the exponential correlation matrix $R_p(\theta(\tau))$ evaluated in $\tau = 0$ for $1 \leq p \leq 2$ is a standard $p$-norm distance matrix corresponding to the point set $\{\tilde{x}^1, \ldots, \tilde{x}^n\} \subset \mathbb{R}^d$, up to the sign. The Gaussian case $p = 2$ differs considerably from the remaining cases $1 \leq p < 2$:

I) For the Gaussian case $p = 2$, the matrices $R'_2(\theta(0)) := d/d\tau R_p(\theta(0))$ are standard Euclidean distance matrices and thus are \textit{rank deficient}. More precisely,

$$\text{rank}(R'_2(\theta(0))) \leq d + 2, \quad \text{for all } n \geq d + 2,$$

see [21, Theorem 2.1]. Here, the term \textquote{Euclidean distance matrix} is used in the sense of [20,21], where the matrix entries are the \textit{squared} Euclidean inter-points distances $(\|\tilde{x}^i - \tilde{x}^j\|_2^2)$. In contrast, Baxter [5] uses the same term for distance matrices of the form $(\|\tilde{x}^i - \tilde{x}^j\|_2^2)_{i,j \leq n}$, which are proven to be regular for distinct sample points.

II) In the non-Gaussian case, there also exist sets of mutually distinct sample points, for which the matrices $-(\|\tilde{x}^i - \tilde{x}^j\|_p^2)_{i,j \leq n}$ are singular, e.g. take $\{(0,0), (1,0), (0,1)\} \subset \mathbb{R}^2$, see [5, Ex. 2.3.4].

Yet, the following proposition shows that singular matrices are excluded, if the $k$th coordinates of the points $\tilde{x}^i$ ($i = 1, \ldots, n$) are pairwise distinct univariate point sets for all $k = 1, \ldots, d$. An equivalent way to express the latter requirement is to say that the set $\{\tilde{x}^1, \ldots, \tilde{x}^n\} \subset \mathbb{R}^d$ of sample points forms a \textit{Latin Hypercube Sampling}, see e.g. [25, §1.4].

Recall that a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called \textit{(strictly) almost negative definite}, iff $v^TAv \leq 0$ (resp. $< 0$) for all vectors in the hyperplane $v \in 1^\perp \subset \mathbb{R}^n$, see [5] and references therein.

**Proposition 1.** Let $1 \leq p < 2$ and let $\{x^1, \ldots, x^n\} \subset \mathbb{R}^d$ be a Latin Hypercube sampling of at least $n \geq 2$ sample points. Then, the $p$-norm distance matrix $D_p = (\|x^i - x^j\|_p^p)_{i,j \leq n} \in \mathbb{R}^{n \times n}$ is strictly almost negative definite and thus regular.

Moreover, $1^T D_p 1 > 0$, where $1 = (1, \ldots, 1)^T \in \mathbb{R}^n$.

**Proof.** For $1 \leq p < 2$, there exists $0 < \delta \leq 1$ such that $p = 2 - \delta$. Let $\bar{p} := 2\sqrt{1 - \delta/2}$, $s := \bar{p}/2$. Then, $\bar{p}s = p$ and

$$|x^i_k - x^j_k|^p = ((|x^i_k - x^j_k|\bar{p})^{1/\bar{p}})^p = (\|x^i_k - x^j_k\|_\bar{p}^\bar{p})^{s} = (\|x^i_k - x^j_k\|_{\bar{p}}^{\bar{p}s}).$$

Because $\bar{p} \in (0, 2)$ and $s \in (0, 1)$, Corollary 2.3.10 of [5] applies to $D^{(k)}_p := (|x^i_k - x^j_k|^p)_{i,j \leq n}$ for $k = 1, \ldots, d$, showing that these matrices are \textit{regular} and in particular \textit{strictly almost negative definite}.

As a consequence, $D_p = \sum_{k=1}^d D^{(k)}_p$ is strictly almost negative definite with zero trace and thus regular by the simple argument of [5, Prop. 2.2.3].
Moreover, let $b^1 := 1/\sqrt{n}1$ and let $b^2, \ldots, b^n$ form an orthonormal basis of $1^\perp$.

Then $Q := (b^1, \ldots, b^n)$ is orthogonal and $0 = \text{tr}(D_p) = \text{tr}(Q^TD_pQ) = (b^1)^TD_pb^1 + \sum_{k=2}^{n}(b^k)^TD_pb^k$. Hence, $(b^1)^TD_pb^1 > 0$, because all other terms in this sum are strictly negative. □

By the same argument, the derivatives of the non-Gaussian exponential correlation matrices based on Latin Hypercube samplings along the coordinate lines $\theta(\tau) = \tau e_k$ are regular in $\tau = 0$ ($k = 1, \ldots, n$).

In combination with Proposition 1, the next proposition shows in particular that the prerequisites of Theorem 2 to be found in the next section are fulfilled for the non-Gaussian exponential correlation models.

**Proposition 2.** Let $A \in \mathbb{R}^{n \times n}$ symmetric be strictly almost negative definite with zero diagonal. Then $1^TA^{-1}1 \neq 0$.

**Proof.** The fact that $A$ is almost negative definite may be expressed by stating that the matrix $B = -(I - 1/n11^T)A(I - 1/n11^T)$ is positive semi-definite, see [20, eq. (1)]. In particular, it holds $B1 = 0$ and $z^TBz > 0$ for all vectors in the hyperplane $z \in (\text{span } 1)^\perp$. Hence, the rank of $B$ is precisely $n - 1$.

Since $A$ is almost negative definite, it is Euclidean, i.e. there exist $n$ vectors $v^1, \ldots, v^n \in \mathbb{R}^r$ in Euclidean space of some dimension $r$, such that $A_{ij} = ||v^i - v^j||^2$. This is, again, a result due to Schoenberg [26], see [5, Theorem 2.2.4] for an elementary proof.

Moreover, the rank of the matrix $B$ is precisely the smallest dimension $r$ for which a distance-generating point set $v^1, \ldots, v^n \in \mathbb{R}^r$ can be found, see [20, §3]. The number $r$ is called the *dimensionality* of $A$ as an EDM. Thus, $A$ has full rank $n$ and its dimensionality is rank($B$) = $n - 1$. But by [20, Theorem 6], this can only be if $1^TA^{-1}1 \neq 0$. (By the Corollary to Theorem 1 in [20, p. 84], indeed $1^TA^{-1}1 < 0$.) □

2.3. *On the differentiability of the eigenvalue decomposition of the exponential correlation matrices in $\theta = 0$*

Let $\theta : \tau \rightarrow \theta(\tau)$ be a smooth curve in $\mathbb{R}^d$. Then, $\tau \mapsto R_p(\theta(\tau))$ is a smooth curve of symmetric matrices in $\mathbb{R}^{n \times n}$. By [27, Theorem 7.6], *all the eigenvalues and eigenvectors of $R_p(\theta(\tau))$ can be chosen smoothly in $\tau$, even in the case of multiple eigenvalues all along $\tau*, provided that no different eigenvalues meet at infinite order, as in the counter example [27, Example 7.7]. This is a generalization of a considerably old result by F. Rellich from 1937 [28, Satz 1] cited in English as *Result 7.2* in [27]. In fact, for the considerations in this work, Rellich’s older result suffices, which assures that the eigenvalue decomposition is real-analytic in $\tau$, provided $\theta(\tau)$ is real analytic.

In particular, being the roots of a polynomial with real analytic coefficients, the eigenvalue functions $\lambda_i(\theta(\tau))$ ($i = 1, \ldots, n$) allow for *Taylor expansions to arbitrary order.*
For similar results from a more computational perspective, see [29]. Here, an algorithm to actually compute eigenvector derivatives is presented. The method works for multiple eigenvalues, provided that for some order \( m \in \mathbb{N} \), the \( m \)-th order derivatives of the eigenvalues are mutually distinct, which prevents the pathological case of different eigenvalues meeting at infinite order. Explicit formulae for the derivatives of multiple eigenvalues have already been featured in [30, Theorems 6–10 and Lemmata 1 & 2].

Perhaps, the aforementioned facts require some remarks in regard of classical matrix perturbation theory. It is a well-known fact that “one cannot expect the eigenvectors of nearby matrices to lie near one another when their corresponding eigenvalues belong to poorly separated eigenvalues” (Stewart [31, p. 728]). This is illustrated in the very introduction of the same reference by an example of \( (3 \times 3) \) matrices \( A'(\epsilon) \), \( A''(\epsilon) \) differing from a diagonal matrix \( A \) with a two-fold eigenvalue by terms of order \( \epsilon \). In the parlance of [27], however, both Stewart’s \( A'(\epsilon) \) and \( A''(\epsilon) \) are real analytic Hermitian curves in \( \epsilon \) and both feature eigenvector bases that pass (after normalizing and fixing of orientation) smoothly through \( \epsilon = 0 \), where the multiple eigenvalue occurs.

3. A theoretical explanation of the Gaussian condition number anomaly

The condition number \( \kappa_2(R) \) of a symmetric positive definite matrix \( R \) with respect to the Euclidean norm may be expressed as the ratio between the maximum and the minimum eigenvalue. Throughout the paper, it is assumed that the eigenvalues of a given (positive definite) correlation matrix are in decreasing order, so that \( \lambda_1 \geq \ldots \geq \lambda_n > 0 \) and \( \kappa_2(R) = \lambda_1/\lambda_n \).

When the hyper-parameters approach zero, then for every admissible choice of parametrized correlation functions, the correlation matrix converges to the singular matrix with every entry equal to 1, i.e. \( \lim_{\|\theta\| \to 0} R(\theta) = 11^T \). Here, boldface \( 1 = (1, \ldots, 1)^T \in \mathbb{R}^n \) denotes the vector with all entries equal to 1. Hence

\[
\kappa_2(R(\theta)) \to \infty, \quad \|\theta\| \to 0.
\]

This shows that all feasibly parametrized correlation matrices become arbitrarily ill-conditioned for hyper-parameters \( \|\theta\| \to 0 \). What distinguishes the Gaussian correlation from the remaining exponential correlation models is how fast the condition number blows up.

In the following, it will be proved that the conditioning of the Gaussian correlation matrix grows at least as fast as \( 1/\|\theta\|^2 \), while it grows generally only as fast as \( 1/\|\theta\| \) for the remaining members of the exponential correlation family along sequences \( \mathbb{R}_{>0} \ni \tau \mapsto \theta(\tau) \) such that \( \theta(0) = 0 \).

**Theorem 1.** Let \( x^1, \ldots, x^n \in \mathbb{R}^d \), be a set of \( n \geq d + 4 \) arbitrary mutually distinct sample points and let \( R(\theta) = (\exp(-\sum_{k=1}^d \theta_k |x^i_k - x^j_k|^2))_{i,j} \in \mathbb{R}^{n \times n} \) be the auto-correlation matrix corresponding to the Gaussian correlation model. Let \( \theta : \tau \mapsto \theta(\tau) \) be a real
analytic curve with $\theta(0) = 0$, $\theta'(0) = h \in \mathbb{R}^d$. Then, for $\tau \to 0$, the condition number $\kappa_2(R(\theta(\tau)))$ grows at least quadratically in $1/\tau$. More precisely, there exists a constant $c > 0$ such that

$$
\kappa_2(R(\theta(\tau))) \geq \frac{c}{\tau^2}, \quad \tau \to 0.
$$

**Proof.** In the setting of Theorem 1, it holds $R(0) = 11^T$. Let $\lambda_1(\theta) \geq \ldots \geq \lambda_n(\theta)$ be the eigenvalues of $R(\theta)$ in descending order. Since the correlation matrix is symmetric, it features a set of mutually orthonormal eigenvectors. Let $q_1(\theta), \ldots, q_n(\theta) \in \mathbb{R}^n$ be such eigenvectors corresponding to the decreasing ordering of the eigenvalues so that the matrix $Q(\theta) = (q_1(\theta), \ldots, q_n(\theta))$ is orthogonal, $Q(\theta)^TQ(\theta) = I_{n \times n}$. Then

$$
R(\theta) = Q(\theta)A(\theta)Q(\theta)^T, \quad A(\theta) = \text{diag}(\lambda_1(\theta), \ldots, \lambda_n(\theta)).
$$

As pointed out in Section 2.3, the eigenvalues and eigenvectors may be chosen to vary analytically along the curve $\theta(\tau)$ (after normalization and fixing of ordering and orientation), even when passing through $\theta(0) = 0$. For convenience, write $R(\tau) := R(\theta(\tau))$ and $R'(\tau) := d/d\tau R(\theta(\tau))$ and so forth.

The theorem will be proved by showing that $\lambda_n'(0) = 0$ and applying Taylor’s Theorem. Since $R(0)1 = (11^T)1 = n1$ and $R(0)v = 0$ for all $v \in \text{span}\{1\}$, it holds $A(0) = \text{diag}(n, 0, \ldots, 0)$ and the vectors $q_1(0) = \pm n^{-1/2}1$ are the two possible choices of normalized eigenvectors corresponding to the single eigenvalue $\lambda_1(0) = n$. Without loss of generality, suppose that the plus sign has been followed along $\|\theta(\tau)\| \to 0$. Moreover, $(q_2(0), \ldots, q_n(0))$ is an orthonormal basis of $\text{span}\{1\}^\perp$. As a consequence,

$$
R'(0) = Q(0)A'(0)Q^T(0) + Q'(0)A(0)Q^T(0) + Q(0)A(0)(Q')^T(0)
$$

$$
= Q(0)A'(0)Q^T(0) + n(q_1'(0)q_1^T(0) + q_1(0)(q_1')^T(0))
$$

$$
= Q(0)A'(0)Q^T(0) + nQ(0)(Q'(0)q_1^T(0) + q_1(0)(q_1')^T(0))Q(0)Q^T(0)
$$

$$
= Q(0)SQ(0)^T,
$$

where

$$
S = \begin{pmatrix}
\lambda_1'(0) & n\langle q'_2(0), q_1(0) \rangle & \ldots & n\langle q'_n(0), q_1(0) \rangle \\
n\langle q'_2(0), q_1(0) \rangle & \lambda_2'(0) & \ldots & \lambda_n'(0) \\
\vdots & \ddots & \ddots & \vdots \\
n\langle q'_n(0), q_1(0) \rangle & \ldots & \lambda_n'(0)
\end{pmatrix}.
$$

This means that the same similarity transformation $Q(0)$ that diagonalizes $R(0)$ brings the derivative $R'(0)$ to arrowhead form. Since $R'(0)$ and the arrow matrix $S$ are identical.
up to the similarity transformation by $Q(0)$, they share the same rank. Because the derivative of the Gaussian correlation matrix by $\tau$ along $\theta(\tau)$ evaluated in $\tau = 0$ is a standard Euclidean distance matrix, it holds

$$\text{rank}(R'(0)) \leq d + 2$$

for all $n \geq d + 2$, see Section 2.2. Hence,

$$d + 2 \geq \text{rank}(R'(0)) = \text{rank}(S) \geq \dim(\text{span}\{Se_j | j = 2, \ldots, n\})$$

$$= \dim \left( \text{span} \left\{ \begin{pmatrix} n\langle q_2(0), q_1(0) \rangle \\ \lambda_j^2(0) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} n\langle q_n(0), q_1(0) \rangle \\ 0 \\ \vdots \\ \lambda_n^2(0) \end{pmatrix} \right\} \right).$$

Therefore, $\lambda_j'(0) = 0$ for all but $d + 2$ exceptions at the most. By Lemma 1 to be found in Appendix A, the possible exceptions $\lambda_j'(0) \neq 0$ correspond to the eigenvalues of smallest index, i.e. $\lambda_2'(0) \neq 0, \ldots, \lambda_{d+3}'(0) \neq 0$ at the most. Hence, $\lambda_j'(0) = 0$ for $j = d + 4, \ldots, n$. In particular, $\lambda_n'(0) = 0$ and $\lambda_j'(0) = 0$ for all $k = 1 \ldots, d$, so that the Taylor expansion of $\lambda_n$ in $\tau = 0$ is

$$\lambda_n(\tau) = \lambda_n(0) + \lambda_n'(0)\tau + \mathcal{O}(\tau^2) = \mathcal{O}(\tau^2).$$

Hence, for $\tau \to 0$ and a suitable constant $c > 0$,

$$\kappa_2(R(\tau)) = \frac{\lambda_1(\tau)}{\lambda_n(\tau)} \geq \frac{c}{\tau^2}. \quad \square$$

It is not claimed that the growth rate estimate established in Theorem 1 is sharp. If higher order derivatives of $\lambda_n(0)$ can be shown to also vanish, then the condition number is proved to grow even faster. The very low rank of $R'(0)$ and the plots in Figs. 1, 2 suggest that this is actually the case for the examples considered in Section 4.

The rank deficiency of the partial derivative matrices occurs for the Gaussian power $p = 2$ and is for arbitrary scattered Latin Hypercube samplings not shared by the other members of the family of exponential auto-correlation functions $\rho(\theta) = \exp(-|x_i - x_j|^p)$, $1 \leq p < 2$, see Section 2.2. It is precisely this property that distinguishes the Gaussian model from the remaining family of exponential correlation functions. The most important consequence of the rank deficiency of the derivatives of the correlation matrix along analytical curves $\tau \mapsto \theta(\tau)$ is that the derivative of the smallest eigenvalue of $R$ vanishes in $\tau = 0$. In this regard, the next theorem may be considered as a reversion of Theorem 1.

Please cite this article in press as: R. Zimmermann, On the condition number anomaly of Gaussian correlation matrices, Linear Algebra Appl. (2014), http://dx.doi.org/10.1016/j.laa.2014.10.038
Theorem 2. Let $R(\theta)$ be an arbitrary correlation matrix corresponding to a point set $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, parametrized by the hyper-parameter vector $\theta \in \mathbb{R}^d_{>0}$ such that $R(\theta)$ is real analytic with respect to $\theta$ and $R(\theta) \to \mathbf{1}\mathbf{1}^T$, for $||\theta|| \to 0$. Let $\theta : \tau \mapsto \theta(\tau)$ be a real analytic curve with $\theta(0) = 0$, $\theta'(0) = \mathbf{h} \in \mathbb{R}^d_{>0}$. If $R'(0) := d/d\tau R(\theta(\tau))$ is regular and if $\mathbf{1}^T R'(0)^{-1} \mathbf{1} \neq 0$, then the derivative of $\lambda_n(\theta(\tau))$ does not vanish in $\tau = 0$. Moreover, the condition number $\kappa_2(R(\theta(\tau)))$ grows at most as fast as $1/\tau$. More precisely, there exists a constant $c > 0$ so that

$$\kappa_2(R(\theta(\tau))) \leq \frac{c}{\tau}.$$ 

Remark. The conditions of the above theorem can be relaxed to hold for smooth curves $\tau \mapsto \theta(\tau)$ along which no two eigenvalues of $R(\theta(\tau))$ meet at infinite order, see [27, §7.6].

Proof. In the setting of Theorem 2, it holds $R(0) = \mathbf{1}\mathbf{1}^T$ and the eigenvector corresponding to the largest eigenvalue $\lambda_1(0)$ is $q_1(0) = n^{-1/2} \mathbf{1}$. It holds $\lambda_2(0) = \ldots = \lambda_n(0) = 0$.

As outlined in Section 2.3, the eigenvalues and eigenvectors pass real analytically through $\theta(0) = 0$.

Let $R'(0)$ be regular. As a first step, we prove that $\lambda_j'(0) \neq 0$ for all $j = 2, \ldots, n$, if and only if $\mathbf{1}^T R'(0)^{-1} \mathbf{1} \neq 0$. Keeping the notation introduced in the proof of Theorem 1, by (5), it holds

$$\mathbf{1}^T R'(0)^{-1} \mathbf{1} = \mathbf{1}^T Q(0) S^{-1} Q(0)^T \mathbf{1} = n e_1^T S^{-1} e_1 = n S_{11}^{-1},$$

where $e_1 \in \mathbb{R}^n$ is the first unit vector. Let $s^1, \ldots, s^n \in \mathbb{R}^n$ denote the columns of the arrow matrix $S$ from (6). By Cramer’s rule,

$$S_{11}^{-1} = \frac{\det(e_1, s^2, \ldots, s^n)}{\det(S)},$$

see any primer on Linear Algebra. Since $R'(0)$ is assumed to be regular, so is $S$ by (5), i.e. $\det(S) \neq 0$. Hence,

$$\mathbf{1}^T (R'(0))^{-1} \mathbf{1} \neq 0 \Leftrightarrow S_{11}^{-1} \neq 0 \Leftrightarrow \det(e_1, s^2, \ldots, s^n) = \prod_{j=2}^n \lambda_j'(0) \neq 0.$$ 

In particular, $\lambda'_n(0) \neq 0$ and by Taylor’s Theorem $\lambda_n(\tau) = \lambda'_n(0) \tau + O(\tau^2)$ with $\lambda'_n(0) \neq 0$. Therefore,

$$\kappa_2(R(\theta(\tau))) = \frac{\lambda_1(\theta(\tau))}{\lambda_n(\theta(\tau))} = \frac{\lambda_1(\theta(\tau))}{\lambda'_n(0) \tau + O(\tau^2)} = O(1/\tau).$$

$\square$
Obviously, the Gaussian correlation model does not fulfill the essential requirement of Theorem 2 that $R'(0)$ be regular, while the other members of the exponential family for $1 \leq p < 2$ do for arbitrary Latin Hypercube sampling plans, as stated in Section 2.2. In general, if $R'(0)$ is regular, then it can be shown that the derivatives of all eigenvalues do not vanish in $\theta(0) = 0$ apart from at most one possible exception, see [12, Lemma A.1]. The additional condition $1^T R'(0)^{-1} 1 \neq 0$ postulated in Theorem 2 precludes this possible exception. The key feature of Theorem 2 is that its conditions may easily be checked for a given sampling plan and any given correlation model by the following simple algorithm without having to compute eigenvalue or eigenvector derivatives:

1. Compute $R'(0) = d/\tau d R(\theta(0))$, check for regularity. (For the family of exponential correlation functions and, say $\tau \mapsto \theta(\tau) = \tau h$, simply $R'(0) = (-\sum_{k=1}^{d} h_k |x_k - x_k'| ^{p})_{i,j} \in \mathbb{R}^{n \times n}$, $1 \leq p < 2$.)

2. Compute the solution $x$ to the linear system $R'(0)x = 1$.

3. Check if $\langle x, 1 \rangle \neq 0$. If so, then the requirements of Theorem 2 hold true and the condition number of the correlation matrix is comparably good-natured for small values of $\tau$.

4. Numerical experiments

In this section, the facts established in Theorems 1 and 2 are illustrated by numerical experiment. To this end, two sets of sample points selected randomly from the interval $[0, 100]$ are considered, each consisting of $n = 10$ points. The precise sample locations are listed in Table 1. First, the univariate case is considered. Then the sample points in Table 1 are combined in order to build a set of two-dimensional sample points, where $x^1 = (58.225, 84.072)^T \in \mathbb{R}^2$ and so forth.

To begin with, the condition numbers of the univariate exponential correlation matrices $R_p(\theta) = \exp(-\theta |x_i - x_j|^p)_{i,j} \in \mathbb{R}^{10 \times 10}$ corresponding to the selected sampling plans are compared for powers $p = 1, 1.5, 1.9, 1.99, 2$. The endmost value $p = 2$ gives the Gaussian correlation. Figs. 1 and 2 display the condition numbers over $0.001 \leq \theta \leq 0.01$ for the various matrices. Both plots reflect exactly the behavior predicted by Theorems 1 and 2. In the standard scale (leftmost boxes of Figs. 1, 2), the condition number plots of the exponential models appear only as flat lines in comparison to the Gaussian condition number graph. Even when compared to the exponential correlation model corresponding to $p = 1.99$, which is numerically close to the Gaussian model ($p = 2$), the conditioning of the Gaussian correlation matrix is approximately as large as the condition number of

---

Please cite this article in press as: R. Zimmermann, On the condition number anomaly of Gaussian correlation matrices, Linear Algebra Appl. (2014), http://dx.doi.org/10.1016/j.laa.2014.10.038
Fig. 1. Condition numbers of the correlation matrices \( R_p(\theta) = \exp(-\theta |x^i - x^j|^p) \) sub \( \in \mathbb{R}^{10 \times 10} \) corresponding to sampling plan (A) from Table 1 for powers \( p = 1, 1.5, 1.9, 1.99, 2 \). Left: standard scale, middle: logarithmic scale, right: standard scale with the graph of the Gaussian condition number omitted.

Fig. 2. Same as Fig. 1 but for sampling plan (B) listed in Table 1.
Fig. 3. Same as Fig. 2, middle, but with the plots of $\theta \mapsto 10/\theta$ (triangle) and $\theta \mapsto (5/\theta)^2$ (square) added.

Fig. 4. Condition numbers for the exponential correlation matrices for a set of two-dimensional sample points, along $\theta(\tau) = \tau(1, 1)^T$ for $\tau \in [10^{-5}, 10^{-4}]$ compared to $O(1/\tau)$ and $O(1/\tau^2)$ with constants adjusted for a well-arranged plot.

the exponential correlation matrix squared. In Fig. 3, the plots of the condition numbers of the exponential and Gaussian correlation matrices based on sampling plan (B) from Table 1 are compared to the functions $\theta \mapsto 10/\theta$ and $\theta \mapsto (5/\theta)^2$. This plot suggests that the growth rate of the condition number of the Gaussian correlation matrix is in fact higher than that of $O(1/\theta^2)$. Finally, Fig. 4 shows the condition number growth for the exponential correlation matrices for the two-dimensional sampling constructed.
from Table 1 along the line \( \theta(\tau) = \tau(1,1)^T \in \mathbb{R}^2 \) for \( \tau \in [10^{-5}, 10^{-4}] \), where the same phenomena as in the univariate case are observed.

A word of caution: The correlation matrix will also become ill-conditioned regardless of the values of the associated hyper-parameters, if there are two sample points very close to each other. In this case, the numerical rank computation may be superposed by round-off errors.

Appendix A

**Lemma 1.** Let \( R(\theta(\tau)) \) be the parametrized Gaussian correlation matrix as in Theorem 1.

Let \( R(\theta(\tau)) = Q(\theta(\tau)) \Lambda(\theta(\tau)) Q(\theta(\tau))^T \) be the parametrized eigenvalue decomposition, where \( \Lambda(\theta(\tau)) = \text{diag}(\lambda_1(\theta(\tau)), \ldots, \lambda_n(\theta(\tau))) \) features the eigenvalues on the diagonal ordered such that \( \lambda_1(\theta(\tau)) \geq \ldots \geq \lambda_n(\theta(\tau)) \) for \( 0 < \tau \leq \varepsilon \). Let \( j, l \in \{2, \ldots, n\} \). If \( \lambda_j'(\theta(0)) = 0 \) and \( \lambda_l'(\theta(0)) \neq 0 \), then \( l < j \).

**Proof.** For convenience, write \( \lambda_j(\tau) := \lambda_j(\theta(\tau)) \). Suppose that the lemma is wrong. Then \( \lambda_j'(0) = 0 \) and \( \lambda_l'(0) \neq 0 \), but \( j < l \). As explained in the proof of Theorem 1, \( \lambda_j(0) = 0 \) and \( \lambda_l(0) = 0 \). Consider the limit behavior \( \tau \to 0 \). As argued in Section 2.3, Taylor’s Theorem applies, leading to

\[
\frac{\lambda_j(\tau)}{\lambda_l(\tau)} = \frac{0.5\lambda_j''(0)\tau^2 + O(\tau^3)}{\lambda_l'(0)\tau + O(\tau^2)} \to 0 \quad (\tau \to 0).
\]

Hence, there exists \( \varepsilon > 0 \) such that \( |\lambda_j(\tau)| < |\lambda_l(\tau)| \), \( 0 < \tau \leq \varepsilon \). Because \( R(\tau) \) is positive definite for \( \tau > 0 \) by the theorem of Schoenberg cited in Section 2.1, it holds \( \lambda_j(\tau) < \lambda_l(\tau) \), in contradiction to the decreasing order of the eigenvalues. \( \square \)

References


Please cite this article in press as: R. Zimmermann, On the condition number anomaly of Gaussian correlation matrices, Linear Algebra Appl. (2014), http://dx.doi.org/10.1016/j.laa.2014.10.038