Maximum Entropy Kernels for System Identification

Francesca P. Carli, ∗ Tianshi Chen, † Lennart Ljung, †

Abstract

A new nonparametric approach for system identification has been recently proposed where the impulse response is modeled as the realization of a zero–mean Gaussian process whose covariance (kernel) has to be estimated from data. In this scheme, quality of the estimates crucially depends on the parametrization of the covariance of the Gaussian process. A family of kernels that have been shown to be particularly effective in the system identification framework is the family of Diagonal/Correlated (DC) kernels. Maximum entropy properties of a related family of kernels, the Tuned/Correlated (TC) kernels, have been recently pointed out in the literature. In this paper we show that maximum entropy properties indeed extends to the whole family of DC kernels. The maximum entropy interpretation can be exploited in conjunction with results on matrix completion problems in the graphical models literature to shed light on the structure of the DC kernel. In particular, we prove that DC kernels admit a closed-form inverse, determinant and factorization. Maximum likelihood properties of the DC kernel are also highlighted. These results can be exploited both to improve the stability and to reduce the computational complexity associated with the computation of DC estimators, as detailed in the paper.

1 Introduction

Most of the currently used techniques for linear system identification relies on parametric prediction error methods (PEMs), Ljung [1999], Soderstrom and Stoica [1989]. Here, finite–dimensional hypothesis spaces of different order, such as ARX, ARMAX or Laguerre models, are first postulated. Then, the most adequate model order is selected trading–off between bias and variance to avoid overfitting. Model–order selection is usually performed by optimizing some penalized goodness–of–fit criteria, such as the Akaike information criterion (AIC) [Akaike 1974] or the Bayesian information criterion (BIC) [Schwarz 1978], or via cross validation (CV) [Hastie et al. 2008]. Statistical properties of prediction error methods are well understood under the assumption that the model class is fixed. Nevertheless, sample properties of PEM approaches equipped e.g. with AIC or CV can much

---

∗Francesca P. Carli is with the Department of Engineering, University of Cambridge, Cambridge, United Kingdom, and with the Department of Electrical Engineering and Computer Science, University of Li`ege, Li`ege, Belgium.
†Tianshi Chen and Lennart Ljung are with the Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, Sweden.
depart from those predicted by standard (i.e. without model selection) statistical theory
(Pillonetto and De Nicolao [2010], Pillonetto et al [2011], Chen et al [2012]).

Motivated by these pitfalls, a new approach to system identification has been recently
proposed where the system impulse response is seen as the realization of a zero–mean
Gaussian process with a suitable covariance that depends on few hyperparameters, learnt
from data via, e.g., marginal likelihood maximization. This procedure can be seen as the
counterpart of model order selection in the parametric paradigm and in many cases it has
been proved to be more robust than AIC-type criteria and CV.

In this scheme, quality of the estimates crucially depends on the covariance (kernel)
of the Gaussian process. A large variety of kernel functions have been introduced in the
machine learning literature Shawe-Taylor and Cristianini [2004]. Nevertheless, a straight
application of standard machine learning kernels in the framework of system identification
is doomed to fail mainly because of the lack of constraints on system stability. For this
reason, several kernels have been recently introduced in the system identification literature
Pillonetto and De Nicolao [2010], Pillonetto et al [2010, 2011], Chen et al [2011, 2012].

This paper focuses on Diagonal/Correlated (DC) kernels, which were introduced by a
deterministic argument in Chen et al [2012], where their performance was assessed on a data
bank of test systems and data sets. In this paper we show that DC kernels admit a maximum
entropy interpretation. Maximum entropy properties of a related family of kernels, the TC
kernels, have been pointed out in the literature, see Carli [2014], Pillonetto and De Nicolao
[2011]. Here we show that arguments in Carli [2014], resting on the theory of matrix
completion problems in the graphical models literature Dempster [1972], Grone et al [1984],
Dym and Gohberg [1981], Gohberg et al [1993], Dahl et al [2008], extend to the whole
family of DC kernels. This approach leads to a closed form expression for the inverse and
determinant of the DC kernel, as well as to a closed form factorization. Maximum likelihood
properties of the DC kernel are also highlighted. We recall that the TC kernel can be seen
both as a particularization of the DC kernel for \( \rho = \sqrt{\lambda} \) Chen et al [2011] (see below) and
as a member of the family of stable spline kernels Pillonetto et al [2010]. It is interesting to
note that, differently to what happens for the family of DC kernels, the maximum entropy
interpretation does not extend to the family of stable spline kernels (e.g. to second–order
stable spline kernels). These results can be exploited both to improve stability and reduce
the computational burden of computational schemes for the evaluation of DC estimators.

The paper is organized as follows. In Section 2 the problem is introduced and Gaussian
process regression via the DC kernel is briefly reviewed. In Section 3 relevant theory of
maximum entropy matrix completion problems is introduced. Section 4 contains our main
results. In Section 5 properties of the DC kernel are exploited to improve efficiency and
robustness of computational schemes for the evaluation of DC estimators. Section 6 ends
the paper.

**Notation.** Let \( \mathcal{S}_n \) denote the set of symmetric matrices of order \( n \). We write \( A \succeq 0 \) (resp. \( A \succ 0 \)) to denote that \( A \) is positive semidefinite (resp. positive definite). Moreover, we
denote by \( I_k \) the identity matrix of order \( k \), and by \( \text{diag} \{ a_1, a_2, \ldots, a_k \} \) the diagonal matrix
with diagonal elements \( \{ a_1, a_2, \ldots, a_k \} \). If \( A \) is a square matrix of order \( n \), for index sets
\( \beta \subseteq \{ 1, \ldots, n \} \) and \( \gamma \subseteq \{ 1, \ldots, n \} \), we denote the submatrix that lies in the rows of \( A \)
indexed by \( \beta \) and the columns indexed by \( \gamma \) as \( A(\beta, \gamma) \). If \( \gamma = \beta \), the submatrix \( A(\beta, \gamma) \) is
abbreviated by \( A(\beta) \). For \( p, q \in \mathbb{N} \) with \( p < q \), the index set \( \{ p, p+1, \ldots, q \} \) is denoted by
2 Linear system identification via Gaussian Process Regression

The system identification problem for linear discrete-time, time-invariant systems is as follows. Consider the measurement model

\[ y(t) = \sum_{k=1}^{\infty} g(k)u(t-k) + v(t), \quad t = 1, \ldots, N \tag{1} \]

where \( t = 1, \ldots, N \) are time indices at which the measured input \( u(t) \) and output \( y(t) \) are collected, \( v(t) \) is the disturbance, for convenience assumed to be a zero mean white Gaussian noise with variance \( \sigma^2 \), and \( g(t), t = 1, 2, \ldots \), denotes the system impulse response. The goal is to estimate \( g(t) \) based on the data \( \{y(t), u(t)\}_{t=1}^{N} \).

Model (1) can be expressed in matrix form as

\[ Y = \Phi^T g + V \tag{2} \]

where \( Y \) and \( V \) are \( N \)-dimensional vector, \( \Phi^T \in \mathbb{R}^{N \times \infty} \) is a matrix whose entries are defined by the system input, the unknown \( u(t) \) having been set to zero for convenience, and \( g = [g(1) \ g(2) \ \cdots]^T \) is the infinite–dimensional vector representing the impulse response. The problem is to estimate \( g \) from \( Y \) and \( \Phi \).

2.1 Gaussian process regression with DC Kernels

Under the framework of Gaussian process regression [Rasmussen and Williams 2006], the impulse response \( g(t) \) is modeled as a discrete-time Gaussian process with zero mean and a suitable covariance function, independent of the disturbance \( v(t) \):

\[ g(t) \sim \mathcal{GP}(0, K(t, t')) \], \( t, t' = 1, 2, \ldots \) \tag{3} \]

where \( K(t, t') \) is often called the kernel (function). The kernel \( K(t, t') \) is parameterized by a vector of hyperparameters \( \eta \) and can be written as \( K(t, t'; \eta) \). In practice, it is impossible to manipulate the infinite impulse response. Since the impulse response of linear stable systems decays exponentially to zero, it is often enough to truncate the infinite impulse response at a certain order \( n \) and consider a finite one instead. In this case, the vector \( g \) in (2) becomes \( g = [g(1) \ g(2) \ \cdots \ g(n)]^T \) and accordingly \( \Phi^T \in \mathbb{R}^{N \times n} \), and in turn \( g \sim \mathcal{N}(0, K(\eta)) \) where \( K(\eta) \) is the \( n \)-dimensional covariance (kernel) matrix.

According to an Empirical Bayes paradigm [Berger 1985; Maritz and Lwin 1989], the hyperparameters can be estimated from the data via marginal likelihood maximization, i.e. by maximizing the marginalization with respect to \( g \) of the joint density of \( Y \) and \( g \)

\[ \hat{\eta} = \arg \min_{\eta} \left\{ \log \det(\Phi^T K(\eta)\Phi + \sigma^2 I_N) + Y^T (\Phi^T K(\eta)\Phi + \sigma^2 I_N)^{-1} Y \right\} \tag{4} \]

Once \( \eta \) is estimated, the impulse response \( g \) can be computed as the minimum variance estimate given \( Y \) and \( \hat{\eta} \), i.e.

\[ \hat{g} := \mathbb{E}[g|Y, \hat{\eta}] = K(\hat{\eta})\Phi \left( \Phi^T K(\hat{\eta})\Phi + \sigma^2 I_N \right)^{-1} Y. \tag{5} \]
Prior information is introduced in the identification process by assigning the covariance \( K(\eta) \). The quality of the estimates crucially depends on this choice as well as on the quality of the estimated \( \hat{\eta} \). Several kernels have been recently introduced in the system identification literature [Pillonetto and De Nicolao 2010, Pillonetto et al. 2010, Chen et al. 2011, 2012]. A class of prior covariances which has been proved to be very effective in the system identification scenario, is the class of Diagonal/Correlated (DC) kernel [Chen et al. 2011, 2012]:

\[
(K_{DC})_{ij} := K_{DC}(i, j; \eta) = c \lambda^{\frac{|i-j|}{2}} \rho^{|i-j|}, \quad i, j = 1, \ldots, n
\]

(6)

where \( \eta = [c \lambda \rho]^T \) with \( c \geq 0 \), \( 0 \leq \lambda < 1 \), \( -1 \leq \rho \leq 1 \). Here \( \lambda \) accounts for the exponential decay of the impulse response, while \( \rho \) describes the correlation between neighboring impulse response coefficients. If \( \rho = \sqrt{\lambda} \), we obtain the so-called Tuned/Correlated (TC) kernel [Chen et al. 2011]

\[
(K_{TC})_{ij} := K_{TC}(i, j; \eta) = c \lambda^{\max(i,j)}, \quad c \geq 0, \quad 0 \leq \lambda < 1.
\]

(7)

which has also been introduced via a stochastic argument under the name of “first–order stable spline (SS) kernel” in [Pillonetto and De Nicolao 2011].

3 Maximum Entropy band extension problem

Covariance extension problems were introduced by A. P. Dempster [1972] and studied by many authors (see e.g. [Grone et al. 1984], [Dym and Gohberg 1981], [Johnson 1990], [Gohberg et al. 1993], [Dahl et al. 2008] and references therein, see also [Carli et al. 2011, 2013] for an extension to the circulant case). In the literature concerning matrix completion problems, it is common practice to describe the pattern of the specified entries of an \( n \times n \) partial symmetric matrix by an undirected graph of \( n \) vertices which has an edge joining vertex \( i \) and vertex \( j \) if and only if the \((i, j)\) entry is specified. If the graph of the specified entries is chordal (i.e., a graph in which every cycle of length greater than three has an edge connecting nonconsecutive nodes), and, in particular, if the specified elements lie on a band centered along the main diagonal, then the maximum entropy covariance extension problem admits a closed form solution in terms of the principal minors of the matrix to be completed (see, e.g., [Lauritzen 1996]). In this section, we briefly review some fundamental results about maximum entropy band extension problems that will be used to prove our main results in Section 4.

3.1 Entropy

Recall that the differential entropy \( H(p) \) of a probability density function \( p \) on \( \mathbb{R}^n \) is defined by

\[
H(p) = - \int_{\mathbb{R}^n} \log(p(x))p(x)dx.
\]

(8)

In case of a zero–mean Gaussian distribution \( p \) with covariance matrix \( \Sigma_n \), we get

\[
H(p) = \frac{1}{2} \log(\det \Sigma_n) + \frac{1}{2} n \left(1 + \log(2\pi)\right).
\]

(9)
3.2 Extension of matrices

We will discuss $n \times n$ matrices $\Sigma(x)$ where only certain elements are given and fixed, and we want to assign the values of the remaining elements $x$ in such a way that $\Sigma(x)$ becomes a positive definite matrix with as large value as possible of $\det \Sigma(x)$. In view of (9) this is called the maximum entropy extension problem.

We shall in particular study the problem that the fixed values are given in a band around the main diagonal of $\Sigma$, i.e. we shall solve the following problem.

**Problem 3.1 (Maximum entropy band extension problem).** Let $\Sigma(x)$ be a $n \times n$ matrix and $i = 1, \ldots, n$, $j = 1, \ldots, n$. For given $m, \sigma_{ij}, |i - j| \leq m$, the maximum entropy band extension problem is

\[
\begin{align*}
\text{minimize} & \quad \{- \log \det \Sigma(x) \mid \Sigma(x) \in S_n\} \\
\text{subject to} & \quad \Sigma(x) \succ 0 \quad (10b)
\end{align*}
\]

\[
\begin{align*}
\Sigma(x)_{ij} = \sigma_{ij}, \quad |i - j| \leq m \quad (10c)
\end{align*}
\]

To stress the dependence on $n$ and $m$ we shall use the notation $\Sigma^{(m)}_n$ and suppress the argument $x$. The problem is a convex optimization problem. We will denote by $x^o$ its optimal value and by $\Sigma^{(m),o}_n \equiv \Sigma^{(m)}_n(x^o)$ the associated optimal extension.

3.3 Basic results for banded matrices

**Theorem 3.1** ([Dempster, 1972, Dym and Gohberg, 1981]). (i) Feasibility: Problem (10) is feasible, namely $\Sigma^{(m)}_n$ admits a positive definite extension if and only if

\[
\begin{bmatrix}
\sigma_{i,i} & \cdots & \sigma_{i,m+i} \\
\vdots & \ddots & \vdots \\
\sigma_{m+i,i} & \cdots & \sigma_{m+i,m+i}
\end{bmatrix} \succ 0, \quad i = 1, \ldots, n - m \quad (11)
\]

(ii) Bandedness: Assume (11) holds. Then (10) admits a unique solution with the additional property that its inverse is banded of bandwidth $m$, namely the $(i, j)$–th entries of $(\Sigma^{(m),o}_n)^{-1}$ are zero for $|i - j| > m$.

The positive definite maximum entropy extension $\Sigma^{(m),o}_n$ is also called central extension of $\Sigma^{(m)}_n$.

Let $\Sigma$ be such that $[\Sigma]_{ij} = \sigma_{ij}, |i - j| \leq m$. Then, it can be shown ([Dempster, 1972, Dahl et al., 2008]) that Problem (10) is equivalent to the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad \log \det \Sigma^{(m)}_n + \text{trace} \left( \Sigma \left( \Sigma^{(m)}_n \right)^{-1} \right) \\
\text{subject to} & \quad \Sigma^{(m)}_n \succ 0 \quad (12b)
\end{align*}
\]

\[
\begin{align*}
\left( \Sigma^{(m)}_n \right)^{-1}_{ij} = 0, \quad |i - j| > m \quad (12c)
\end{align*}
\]

If we denote with $\theta = [\theta_1, \ldots, \theta_n]^\top$ a zero–mean Gaussian random vector with covariance $\Sigma^{(m)}_n$, then (12c) holds if and only if the random variables $\theta_i, \theta_j$ in $\theta$ are conditionally
independent given the others (see e.g. [Dempster, 1972]). In other words, if we denote with $\Sigma$ the sample covariance of $\theta$, the equivalence between Problem (10) and Problem (12) states that the covariance matrix that maximizes the entropy among all the covariance matrices with given $m$ covariance lags, is also the one that maximizes the likelihood among all the covariance matrices satisfying the conditional independence constraints (12c).

For banded sparsity pattern like those considered so far, Problem (10) admits a closed form solution that can be computed recursively in the following way. We start by considering a partially specified $n \times n$ symmetric matrix of bandwidth $(n - 2)$

$$\Sigma^{(n-2)}_n = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n-1} & x \\ \sigma_{1,2} & \sigma_{2,2} & \cdots & \sigma_{2,n-1} & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{1,n-1} & \sigma_{2,n-1} & \cdots & \sigma_{n-1,n-1} & \sigma_{n-1,n} \\ x & \sigma_{2,n} & \cdots & \sigma_{n-1,n} & \sigma_{n,n} \end{bmatrix}$$

(13)

and consider the submatrix

$$L = [\sigma_{ij}]_{i,j=1}^{n-1}. \quad (14)$$

We call one-step extensions the extensions of $n \times n$, $(n - 2)$–band matrices. The next theorem gives a recursive algorithm to compute the extension of partially specified matrices of generic bandwidth $m$ by computing the one-step extensions of suitable submatrices. It also gives a representation of the solution in factored form.

**Theorem 3.2** ([Gohberg et al., 1993], [Dym and Gohberg, 1981]). (i) The one-step central extension of $\Sigma^{(n-2)}_n$ is given by

$$x^0 = -\frac{1}{y_1} \sum_{j=2}^{n-1} \sigma_{nj}y_j \quad (15)$$

with

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{bmatrix} = L^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (16)$$

Let $\Sigma^{(m)}_n$ be an $n \times n$ partially specified $m$–band matrix. The central extension $C = [c_{ij}]_{i,j=1}^n$ of $\Sigma^{(m)}_n$ is such that for all $m+1 < t \leq n$ and $1 \leq s \leq t-m-1$ the submatrices $C(s : t)$ are the central one-step extensions of the corresponding $(t-s-1)$–band matrix.

(ii) In particular, the central extension of the partially specified symmetric $m$-band matrix $\Sigma^{(m)}_n$ admits the factorization

$$C = \left(L^{(m)}_n U^{(m)}_n \right)^{-1} \quad (17)$$

where $L^{(m)}_n = [\ell_{ij}]$ is a lower triangular banded matrix with ones on the main diagonal, $\ell_{jj} = 1$, for $j = 1, \ldots, n$, and

$$\begin{bmatrix} \ell_{\alpha j} \\ \ell_{\beta j} \end{bmatrix} = -\begin{bmatrix} \sigma_{\alpha \alpha} & \cdots & \sigma_{\alpha \beta} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta \alpha} & \cdots & \sigma_{\beta \beta} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{\alpha j} \\ \vdots \\ \sigma_{\beta j} \end{bmatrix}. \quad (18)$$
for \( j = 1, \ldots, n - 1 \), \( U_n^{(m)} = (I_n^{(m)})^\top \) and \( V = [v_{ij}] \) diagonal with entries

\[
v_{jj} = \left( \begin{bmatrix} \sigma_{jj} & \cdots & \sigma_{j\beta} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta j} & \cdots & \sigma_{\beta\beta} \end{bmatrix} \right)_{1,1}^{-1},
\]

for \( j = 1, \ldots, n \), where

\[
\alpha = \alpha(j) = j + 1 \quad \text{for} \quad j = 1, \ldots, n - 1,
\]

\[
\beta = \beta(j) = \min(j + m, n) \quad \text{for} \quad j = 1, \ldots, n.
\]

### 4 Maximum Entropy properties of the DC kernel

In this section, we provide a proof of the maximum entropy property of the DC kernel that relies on the theory of matrix extension problems introduced so far. This argument leads to a closed form expression for the inverse of the DC kernel as well as to a new factorization where the factors admit a closed–form representation in function of the kernel hyperparameters.

**Proposition 4.1.** Consider Problem \([10]\) with \( m = 1 \) and

\[
\sigma_{ij} = (\mathcal{K}_{DC})_{ij} = \rho^{|i-j|} \lambda^{-\frac{|i-j|}{2}}, \quad |i-j| \leq 1
\]

i.e. consider the partially specified 1–band matrix

\[
\Sigma^{(1)}_n(x) = \begin{bmatrix}
\lambda & \lambda^2 \rho & x_{13} & \cdots & \cdots & x_{1n} \\
\lambda^2 \rho & \lambda^4 \rho & x_{24} & \cdots & \cdots & x_{2n} \\
x_{13} & \lambda^3 \rho & x_{24} & \cdots & \cdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \cdots & \ddots \\
x_{1n} & x_{2n-1} & x_{n-2,1} & \cdots & \cdots & \cdots \\
\end{bmatrix}
\]

Then \( \Sigma^{(1)}_n(x) = \mathcal{K}_{DC} \), i.e. the solution of the Maximum Entropy Problem \([10]\) coincides with the DC kernel \([6]\).

**Proof.** By Theorem \([3,2]\) the maximum entropy completion of \( \Sigma^{(1)}_n(x) \) can be recursively computed starting from the maximum entropy completions of the nested principal submatrices of smaller size. The statement can thus be proved by induction on the dimension \( n \) of the completion.

- Let \( n = 3 \), then by \([15]\)–\([16]\), the central extension of

\[
\begin{bmatrix}
\lambda & \lambda^2 \rho & x_{13} \\
\lambda^2 \rho & \lambda^4 \rho & \lambda^3 \rho \\
x_{13} & \lambda^3 \rho & \lambda^5 \rho \\
\end{bmatrix}
\]

is given by \( x_{13} = \lambda^2 \rho^2 = \mathcal{K}_{DC}(1,3) \), as claimed.
Now assume that the statement holds for $n = k, k \geq 3$, i.e. that $K_{DC}(\{1, \ldots, k\})$ is the central extension of $\Sigma_n^{(1)}(\{1, \ldots, k\})$. We want to prove that $K_{DC}(\{1, \ldots, k+1\})$ is the central extension of $\Sigma_n^{(1)}(\{1, \ldots, k+1\})$. To this aim, we only need to prove that the $(k-1)$ submatrices $K_{DC}(\{s, \ldots, k+1\})$, $1 \leq s \leq k-1$, are the central one-step extensions of the corresponding $(k-s)$-band matrices

$$
\begin{bmatrix}
\lambda^s & \cdots & \lambda^{s+k\rho^{k-s}} & x_{s,k+1} \\
\vdots & \ddots & \vdots & \vdots \\
x_{s,k+1} \lambda^{s+k\rho^{k-s}} & \cdots & \lambda^{k+1}
\end{bmatrix},
$$

or, equivalently, that $x_{s,k+1}^o = K_{DC}(s, k+1)$, for $s = 1, \ldots, k-1$. In order to find $x_{s,k+1}^o$, we consider (16), which, by the inductive hypothesis, becomes

$$
\begin{bmatrix}
y_1^{(s,k+1)} \\
y_2^{(s,k+1)} \\
\vdots \\
y_k^{(s,k+1)} \\
y_{k-s+1}^{(s,k+1)}
\end{bmatrix} = K_{DC}(\{s, \ldots, k\})^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
$$

By considering the adjoint of $K_{DC}(\{s, \ldots, k\})$ one can see that $y_2^{(s,k+1)} = -\frac{\rho}{\sqrt{\lambda}} y_1^{(s,k+1)}$ while all the others $y_i^{(s,k+1)}$, $i = 3, \ldots, k$ are identically zero. It follows that

$$
x_{s,k+1}^o = -\frac{1}{y_1^{(s,k+1)}} y_2^{(s,k+1)} \frac{\lambda^{s+k\rho^{k-s}}}{\sqrt{\lambda}} = \frac{\rho}{\sqrt{\lambda}} \lambda^{s+k\rho^{k-s}} = \lambda^{s+k+1} \rho^{k-s+1} = K_{DC}(s, k+1).
$$

From the equivalence between the maximum entropy problem (10) and the maximum likelihood problem (12) we get the following.

**Proposition 4.2.** Let $m = 1$ and $[\Sigma]_{ij} = [\theta \theta^\top]_{ij} = \lambda^{i-j} \rho^{\|i-j\|}$, $|i-j| \leq 1$, then the DC kernel maximizes the likelihood in (12a) among all covariances that satisfies (12b) and the conditional independence constraints (12c).

**Proposition 4.3.** The following are equivalent

(i) $K_{DC}$ solves Problem (10) with $m = 1$ and moment constraints as in (20).

(ii) $K_{DC}$ admits the factorization

$$
K_{DC} = UWU^\top
$$

(21)
with $U$ Toeplitz and given by

$$U = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i > j \\ \frac{\rho^{i-j}}{\lambda^{j-i/2}} & \text{otherwise} \end{cases} \quad (22)$$

and $W$ diagonal given by

$$W = (1 - \rho^2) \text{diag} \left\{ \lambda, \lambda^2, \ldots, \lambda^{n-1}, \frac{\lambda^n}{1 - \rho^2} \right\} . \quad (23)$$

(iii) $K_{DC}^{-1}$ is tridiagonal banded and is given by

$$(K_{DC}^{-1})_{i,j} = \frac{c_{ij}}{1 - \rho^2} (-1)^{i+j} \lambda^{\frac{i+j}{2}} \rho^{|i-j|} \quad (24)$$

where

$$c_{ij} = \begin{cases} 0 & \text{if } |i - j| > 1, \\ 1 + \rho^2 & \text{if } i = j = 2, \ldots, n - 1, \\ 1 & \text{otherwise.} \end{cases} \quad (25)$$

**Proof.** That $K_{DC}$ admits the factorization $(21)$–$(23)$ follows from Theorem 3.2 (ii). In fact, by $(17)$–$(19)$ the inverse of the DC kernel can be factored as

$$K_{DC}^{-1} = L_n^{(1)} V U_n^{(1)} \quad (26)$$

where $L_n^{(1)}$ takes the form

$$L_n^{(1)} = \begin{cases} 1 & \text{if } i = j \\ \frac{\rho}{\sqrt{\lambda}} & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

and

$$V = \frac{1}{(1 - \rho^2)} \text{diag} \left\{ \frac{1}{\lambda}, \frac{1}{\lambda^2}, \ldots, \frac{1}{\lambda^{n-1}}, \frac{1 - \rho^2}{\lambda^n} \right\} \quad (28)$$

Bandedness of $K_{DC}^{-1}$ follows from Theorem 3.1 (ii) and expression $(24)$ is an immediate consequence of the factorization $(21)$–$(23)$.

**Corollary 4.1.** The DC kernel has determinant

$$\det (K_{DC}) = \lambda^{\frac{n(n+1)}{2}} (1 - \rho^2)^{n-1} . \quad (29)$$

**Proof.** The first (resp., third) factor in the right hand side of $(26)$ is a lower (resp., upper) triangular matrix with diagonal entries equal to one, and hence the positive definite matrix $K_n^{-1}$ and $V$ have the same determinant. □
Remark 4.1. Since TC kernel is a special case of the DC kernel with \( \rho = \sqrt{\lambda} \), the closed-form expressions in [Carli, 2014] for the factors \( U \) and \( W \) of the TC kernel, as well as for its determinant and inverse, can be obtained by setting \( \rho = \sqrt{\lambda} \) in (22), (23), (24) and (29). In particular, we get

\[
U = \begin{cases} 
1 & \text{if } i \leq j \\
0 & \text{otherwise}
\end{cases}
\]

and

\[
W = (\lambda - \lambda^2) \text{diag} \left\{ 1, \lambda, \lambda^2, \ldots, \lambda^{n-2}, \frac{\lambda^{n-1}}{1-\lambda} \right\}.
\]

Note that, whatever the value of the parameter \( \lambda \), all TC kernels share the same factor \( U \).

The inverse of the TC kernel is given by

\[
\frac{1}{\lambda - \lambda^2} \begin{bmatrix} 
1 & -1 & 0 & \ldots & 0 \\
-1 & 1 + \frac{1}{\lambda} & -\frac{1}{\lambda} & \ddots & \vdots \\
0 & \frac{1}{\lambda} & \frac{1}{\lambda} + \frac{1}{\lambda^2} & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & -\frac{1}{\lambda^{n-2}} \\
0 & \ldots & 0 & -\frac{1}{\lambda^{n-2}} & \frac{1}{\lambda^{n-1}} + \frac{1}{\lambda^{n+1}}
\end{bmatrix}
\]

and its determinant is

\[
\det(K_{TC}) = (1 - \lambda)^{n-1} \lambda^{n(n+1)}.
\]

As mentioned above, the TC kernel is a particular stable spline kernel (stable spline kernel of order one – see [Pillonetto et al., 2010]). One may wonder if the maximum entropy interpretation holds for the family of stable spline kernels or not. Unfortunately, it is straightforward to verify that unlike the family of DC kernels, the maximum entropy interpretation does not extend to the family of stable spline kernels (e.g. to second–order stable spline kernels), that do not admit a tridiagonal inverse.

5 Exploiting the DC kernel structure for marginal likelihood maximization

Kernel matrices for system identification are characterized by exponentially decaying main diagonals, so they are often poorly conditioned (see [Chen and Ljung, 2013, Fig. 1] for details). It follows that extra attention has to be paid when computing the cost function in (4).

By noting \( \Phi \in \mathbb{R}^{n \times N} \) and Sylvester’s determinant theorem and matrix inversion lemma, the two terms in (4) can be rewritten as

\[
\det(\sigma^2 I_N + \Phi^\top K_{DC} \Phi) = (\sigma^2)^{N-n} \det(K_{DC}) \det(\sigma^2 K_{DC}^{-1} + \Phi \Phi^\top)
\]

\[
Y^\top (\sigma^2 I_N + \Phi^\top K_{DC} \Phi)^{-1} Y = \frac{Y^\top Y}{\sigma^2} - Y^\top \Phi^\top (\sigma^2 K_{DC}^{-1} + \Phi \Phi^\top)^{-1} \Phi^\top Y
\]

To avoid the computation of \( K_{DC}^{-1} \), we made use of the Cholesky factorization of \( K_{DC} \) and developed a matrix inversion free algorithm [Chen and Ljung, 2013] for accurate and efficient computation of the cost function of (4). In this section, we show that the structure of the DC kernel can indeed be exploited to provide a more stable and efficient algorithm.
Before proceeding to the details, we recall the definition of thin QR factorization cf. [Golub and Van Loan, 1996, p. 230]: if $B = CD$ is a QR factorization of $B \in \mathbb{R}^{p \times q}$ with $p \geq q$, then $B = C(1:p, 1:q)D(1:q)$ is referred to as the thin QR factorization of $B$. We then make two assumptions to guarantee the uniqueness of the thin QR factorization in the following, cf. [Golub and Van Loan, 1996, p. 230, Thm 5.2.2].

**Assumption 5.1.** Assume $\text{rank}[\Phi^\top Y] = n + 1$, and moreover all upper triangular matrices involved in the thin QR factorizations below have positive diagonal entries.

### 5.1 Making use of the structure of DC kernel

From Proposition 4.3, it is easy to see that $\mathcal{K}^{-1}_{DC}$ is tridiagonal and its Cholesky factorization is bidiagonal and has closed-form expression in terms of hyper-parameters:

$$\mathcal{K}^{-1}_{DC} = D D^\top$$  \hfill (33)

where $D$ is an $n$-dimensional lower bidiagonal matrix:

$$D(i, j) = \begin{cases} 
\lambda^{-i/2}(1-\rho^2)^{-1/2}, & \text{for } i = j, j < n \\
-\lambda^{-i/2}(1-\rho^2)^{-1/2}\lambda^{-1/2}/\rho, & \text{for } i = j + 1, j < n \\
\lambda^{-n/2} & \text{for } i = j = n 
\end{cases}$$

Now consider the thin QR factorization of

$$\begin{bmatrix} \Phi^\top & Y \\ \sigma D^\top & 0 \end{bmatrix} = QR = Q \begin{bmatrix} R_1 & R_2 \\ 0 & r \end{bmatrix}$$  \hfill (34)

where $Q$ is an $(N + n) \times (n + 1)$ matrix whose columns are orthogonal unit vectors such that $Q^\top Q = I_{n+1}$, and $R$ is an $(n + 1) \times (n + 1)$ upper triangular matrix. Here, $R$ is further partitioned into $2 \times 2$ blocks with $R_1, R_2$ and $r$ being an $n \times n$ matrix, an $n \times 1$ vector and a scalar, respectively.

Noting $Q^\top Q = I_{n+1}$ yields

$$\sigma^2 \mathcal{K}^{-1}_{DC} + \Phi \Phi^\top = R_1^\top R_1, \Phi Y = R_1^\top R_2, Y^\top Y = R_2^\top R_2 + r^2$$

which from (32) yields that

$$\det(\sigma^2 I_N + \Phi^\top \mathcal{K}_{DC} \Phi) = (\sigma^2)^{N-n} \det(\mathcal{K}_{DC}) \det(R_1)^2$$

$$Y^\top (\sigma^2 I_N + \Phi^\top \mathcal{K}_{DC} \Phi)^{-1} Y = r^2/\sigma^2$$

Further noting (29), the cost function of (4) can be calculated as

$$\frac{r^2}{\sigma^2} + (N - n) \log \sigma^2 + \frac{n(n+1)}{2} \log \lambda + (n - 1) \log(1-\rho^2) + 2 \log \det(R_1)$$  \hfill (35)
5.2 Preprocessing the data

Now we compute the cost function of (4) according to (35). Apparently, the major computation cost relies on the QR factorization (34), which like Chen and Ljung [2013] can be handled in a more efficient way. To see this, consider the thin QR factorization of the matrix

\[
\begin{bmatrix}
\Phi_N^\top & Y_N
\end{bmatrix} = Q_d \begin{bmatrix} R_{d1} & R_{d2} \\
\end{bmatrix}
\]  

(36)

where \(Q_d\) is an \(N \times (n + 1)\) matrix whose columns are orthogonal unit vectors such that \(Q_d^\top Q_d = I_{n+1}\), \(R_{d1}\) is an \((n + 1) \times n\) matrix and \(R_{d2}\) is an \((n + 1) \times 1\) vector.

Now consider further the thin QR factorization of

\[
\begin{bmatrix} R_{d1} & R_{d2} \\
\sigma D^\top & 0
\end{bmatrix} = Q_c R_c
\]  

(37)

where \(Q_c\) is an \((2n + 1) \times (n + 1)\) matrix whose columns are orthogonal unit vectors such that \(Q_c^\top Q_c = I_{n+1}\) and \(R_c\) is an \((n + 1) \times (n + 1)\) upper triangular matrix.

Then from (36) and (37), we have

\[
\begin{bmatrix}
\Phi_N^\top & Y_N
\end{bmatrix} = \begin{bmatrix} Q_d & 0 \\
0 & I_n
\end{bmatrix} \begin{bmatrix} R_{d1} & R_{d2} \\
\sigma D^\top & 0
\end{bmatrix} \begin{bmatrix} Q_c & 0 \\
0 & I_n
\end{bmatrix} = Q_c R_c
\]  

(38)

Noting Assumptions 5.1, comparing (34) with (38) and invoking the uniqueness of the thin QR factorization [Golub and Van Loan, 1996, p. 230, Thm 5.2.2] yields that \(R\) and \(Q\) in (34) can be computed as

\[
R = R_c, \quad Q = \begin{bmatrix} Q_d & 0 \\
0 & I_n
\end{bmatrix} Q_c
\]  

(39)

In this way (using (37) and (39)), we find a more efficient way to compute the QR factorization (34). We thus get the following proposition and algorithm to compute the cost function of (4).

5.3 Proposed algorithm

**Proposition 5.1.** Consider (4). Suppose Assumption 5.1 holds. Then the thin QR factorization (34) can be computed from (37) and (39) and in turn, the cost function of (4) can be computed from (35). Moreover, the regularized impulse response estimate (5) for a given \(\alpha\) can be computed according to

\[
\hat{g} = (\sigma^2 K_{DC}^{-1} + \Phi \Phi^\top)^{-1} \Phi Y = R_1^{-1} R_2
\]  

(40)

**Algorithm 5.1.** Assume that the thin QR factorization (36) has been computed beforehand, and moreover, the matrix \(D\) in (33) is created at each iteration. Then the algorithm consists of the two steps to compute the cost function of (4):

1) compute the QR factorization (37);
2) compute (35).
Now we compare Algorithm 5.1 with Algorithm 2 in [Chen and Ljung, 2013] in terms of computational complexity. Both algorithms consist of three parts:

- **preprocessing:**
  Both Algorithm 5.1 and [Chen and Ljung, 2013, Algorithm 2] require to compute (36) beforehand, which requires $2(n+1)^2(N-(n+1)/3)$ flops according to [Golub and Van Loan, 1996]. This part is done once in the solution of the marginal likelihood maximization.

- **preparation for the evaluation of (4):**
  Algorithm 5.1 needs to create $D^\top$, i.e., the Cholesky factorization of $K_{DC}^{-1}$ and [Chen and Ljung, 2013, Algorithm 2] requires to create $K_{DC}$. Since $D^\top$ is bidiagonal, only $2n-1$ scalars have to be created for $D^\top$. In contrast, since $K_{DC}$ is symmetric, $n(n+1)/2$ scalars are created for $K_{DC}$. This part is done at every iteration in the solution of the marginal likelihood maximization.

- **evaluation of (4):**
  For Algorithm 5.1, straightforward computation of the QR factorization (37) in step 1) requires $2(n+1)^2(2n+1-(n+1)/3)$ flops, and step 2) requires $n+20$ flops. For [Chen and Ljung, 2013, Algorithm 2], the steps 1) and 4), according to Hunger [2007], require $n^3/3+n^2/2+n/6$ and $2n+6$ flops, respectively. The step 2) requires $n^2(n+1)$ flops and the step 3) with straightforward computation of the QR factorization requires $2(n+1)^2(2n+1-(n+1)/3)$ flops according to [Golub and Van Loan, 1996]. This part is done at every iteration in the solution of the marginal likelihood maximization.

To summarize, for the preprocessing part, both the algorithms require the same amount of computations. For the preparation part, for the evaluation of (4) with large $n$, the storage and computational complexity of Algorithm 5.1 is negligible if compared to the one of Algorithm 2 in [Chen and Ljung, 2013]. In particular, for the evaluation of (4) with large $n$, Algorithm 5.1 saves approximately 28% of the computations required by Algorithm 2 in [Chen and Ljung, 2013].

Moreover, we observe that Algorithm 5.1 is more stable than Algorithm 2 in [Chen and Ljung, 2013]. This is because Algorithm 5.1 depends on $D$, which has closed-form expression and thus can be computed exactly, while Algorithm 2 in [Chen and Ljung, 2013] depends on the Cholesky factorization of $K_{DC}$, whose computation can be a problem due to the ill-conditioning of $K_{DC}$. For example, the function `chol` fails to return the Cholesky factor for $K_{DC}$ with $\lambda = 0.6, \rho = 0.95$ and $n = 2000$, in Matlab 2013b.

**Remark 5.1.** Very often, the gradient and Hessian are required to speed up the convergence of the marginal likelihood maximization, and can be computed as follows:

$$\frac{\partial l(K_{DC})}{\partial \eta_i} = \text{trace}((X_1 - X_2) \frac{\partial K_{DC}}{\partial \eta_i})$$

$$\frac{\partial^2 l(K_{DC})}{\partial \eta_i \eta_j} = \text{trace}((X_1 - X_2) \frac{\partial^2 K_{DC}}{\partial \eta_i \eta_j} + (X_1 - X_2) \frac{\partial K_{DC}}{\partial \eta_i} X_2 - (X_1 - X_2) \frac{\partial K_{DC}}{\partial \eta_j} X_1)$$

where $l(K_{DC})$ denotes the cost function of (4), $\eta_i$ denotes the $i$th element of $\eta$, and

$$X_1 = K_{DC}^{-1} - \sigma^2 K_{DC}^{-1} (R_1^T R_1)^{-1} K_{DC}^{-1}$$
$$X_2 = K_{DC}^{-1} R_1^{-1} R_1 R_2 R_2^T K_{DC}^{-1}$$
Finally, note that $K_{DC}^{-1}$ is tridiagonal, the computation of $X_1$ and $X_2$ could be made more efficient by making use of this structure.

**Remark 5.2.** Noting [Chen and Ljung, 2013, Remark 4.2], it is possible to adapt [Chen and Ljung, 2013, Algorithm 2] by using the factorization $K_{DC} = LL^\top$ with $L = UW^{1/2}$ from (21) where $W^{1/2}$ is the square root of $W$. The adapted algorithm is also more stable because it is free of the computation of Cholesky factorization. It only saves approximately 7% computation of [Chen and Ljung, 2013, Algorithm 2] for the evaluation of $[4]$ with large $n$.

### 6 Conclusions

Gaussian process regression for linear system identification has become popular recently, mainly due to the introduction of families of kernels which encode structural properties of the dynamical system. A family of kernels that has been shown to be particularly effective is the family of Diagonal/Correlated (DC) kernels. The main contribution of the present paper is to show that the family of DC kernels admits a maximum entropy interpretation. This interpretation can be exploited in conjunction with results on matrix completion problems in the graphical models literature to shed light on the structure of the DC kernel. In particular, we proved that the DC kernel admits a closed-form inverse, determinant and factorization. Maximum likelihood properties of the DC kernel were also highlighted. The DC kernel structure has been exploited to improve stability and reduce computational complexity associated with the computation of the associated estimator.

### References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Verlag, 1985.

F. P. Carli. On the maximum entropy property of the first–order stable spline kernel and its implications. In *2014 IEEE Multi-conference on Systems and Control*. IEEE, 2014.

F.P. Carli, A. Ferrante, M. Pavon, and G. Picci. A maximum entropy solution of the covariance extension problem for reciprocal processes. *IEEE Transactions on Automatic Control*, 56(9):1999–2012, 2011.

F.P. Carli, A. Ferrante, M. Pavon, and G. Picci. An efficient algorithm for maximum entropy extension of block-circulant covariance matrices. *Linear Algebra and its Applications*, 439 (8):2309–2329, 2013.

T. Chen and L. Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7):2213–2220, 2013.

T. Chen, H. Ohlsson, G. C. Goodwin, and L. Ljung. Kernel selection in linear system identification part II: A classical perspective. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 4326–4331. IEEE, 2011.
T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, 48(8):1525–1535, 2012.

J. Dahl, L. Vanderberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software*, 23:501–520, 2008.

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

H. Dym and I. Gohberg. Extensions of band matrices with band inverses. *Linear algebra and its applications*, 36:1–24, 1981.

I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Classes of Linear Operators, II, OT63*. Birkhäuser, Basel, 1993.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, 1996.

R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra and Its Applications*, 58:109–124, 1984.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2008.

R. Hunger. Floating point operations in matrix-vector calculus. *Tech. Rep. TUM-LNS-TR-05-05, Munich University of Technology*, 2007.

S.L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

L. Ljung. *System Identification - Theory For the User*. Prentice Hall, 1999.

J. S. Maritz and T. Lwin. *Empirical Bayes methods*. Chapman and Hall London, 1989.

G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46:81–93, 2010.

G. Pillonetto and G. De Nicolao. Kernel selection in linear system identification Part I: A Gaussian process perspective. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 4318–4325. IEEE, 2011.

G. Pillonetto, A. Chiuso, and G. De Nicolao. Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *American Control Conference (ACC), 2010*, pages 498–503. IEEE, 2010.

G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: A nonparametric Gaussian regression approach. *Automatica*, 47:291–305, 2011.

C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.
G. Schwarz. Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis. Cambridge University Press, 2004.

T. Soderstrom and P. Stoica. System Identification. Prentice Hall, 1989.