

# Toward Bayesian Data Compression

Johannes Harth-Kitzerow,<sup>\*</sup> Reimar H. Leike, Philipp Arras, and Torsten A. Enßlin

In order to handle large datasets omnipresent in modern science, efficient compression algorithms are necessary. Here, a Bayesian data compression (BDC) algorithm that adapts to the specific measurement situation is derived in the context of signal reconstruction. BDC compresses a dataset under conservation of its posterior structure with minimal information loss given the prior knowledge on the signal, the quantity of interest. Its basic form is valid for Gaussian priors and likelihoods. For constant noise standard deviation, basic BDC becomes equivalent to a Bayesian analog of principal component analysis. Using metric Gaussian variational inference, BDC generalizes to non-linear settings. In its current form, BDC requires the storage of effective instrument response functions for the compressed data and corresponding noise encoding the posterior covariance structure. Their memory demand counteract the compression gain. In order to improve this, sparsity of the compressed responses can be obtained by separating the data into patches and compressing them separately. The applicability of BDC is demonstrated by applying it to synthetic data and radio astronomical data. Still the algorithm needs further improvement as the computation time of the compression and subsequent inference exceeds the time of the inference with the original data.

processed, and analysed. They often reach the limit of the available computational power and storage. Examples include urban technology,<sup>[1]</sup> internet searches,<sup>[2]</sup> bio-informatics<sup>[3]</sup> and radio astronomy.<sup>[4]</sup> In this paper we discuss, how such huge datasets can be handled efficiently by compression.

In general, there are two categories of data compression methods: Lossless compression and lossy compression. From lossless compressed data one can regain the full uncompressed data. This limits the amount of compression as only redundant information can be taken away by a lossless scheme. Lossy compression is more effective in terms of reducing the storage needed by the compressed data. This is possible at the cost of loss of information.

In this work, we focus on lossy compression methods. The considered scenario is: compressing data which carries information about some quantity of

## 1. Introduction

One of the challenges in contemporary signal processing is dealing with large datasets. Those datasets need to be stored,

interest, which we call the signal. Only the relevant information for this signal needs to be conserved. Therefore, there is no need to regain the full original data in such applications.

Many lossy compression schemes have been developed: Rate distortion theory (ref. [5], pp. 301–307) gives a general approach stating the need of a loss function, which shall be minimized in order to find the best compressed representation of some original data  $d$ . As a consequence of the Karhunen–Loève theorem,<sup>[6–8]</sup> principal component analysis (PCA)<sup>[9,10]</sup> can also be used for data compression. Its aim is to compress some data, such that the compressed data carries the same statistic properties as the original. It was shown that PCA minimizes an upper bound of the mutual information of the original and the compressed data about some relevant signal.<sup>[11]</sup> Both methods aim to reproduce the original data from the compressed data, but are not specifically optimized to recover information about the actual quantity of interest.

Before compressing data, one should be clear about the signal on which one wants to keep as much information as possible. In a Bayesian setting, this means that the posterior probability of the signal conditional to the compressed data should be as close as possible to the original posterior that was conditioned on the original data.

The natural distance measure between the original and compressed posterior to be used as the action principle is the Kullback Leibler (KL) divergence.<sup>[12]</sup> From this, we derive Bayesian data compression (BDC). Using the KL divergence as the loss function reduces the problem of finding the compressed

J. Harth-Kitzerow, R. H. Leike, P. Arras, T. A. Enßlin  
Max-Planck-Institut für Astrophysik  
Karl-Schwarzschild-Str. 1, 85748 Garching, Germany  
E-mail: jharthki@mpa-garching.mpg.de

J. Harth-Kitzerow, P. Arras, T. A. Enßlin  
Ludwig-Maximilians-Universität München  
Geschwister-Scholl-Platz 1, 80539 Munich, Germany

J. Harth-Kitzerow  
Technische Universität München  
James-Franck-Str. 1, 85748 Garching, Germany

P. Arras  
Technische Universität München  
Boltzmannstr. 3, 85748 Garching, Germany

J. Harth-Kitzerow, T. A. Enßlin  
Exzellenzcluster ORIGINS  
Boltzmannstr. 2, 85748 Garching, Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/andp.202000508>

© 2021 The Authors. Annalen der Physik published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

DOI: 10.1002/andp.202000508

data representation to an eigenvalue problem equivalent to the generalized eigenvalue problem found by ref. [13]. In this work, we give a didactic derivation and show how this approach can be extended to nonlinear and non-Gaussian measurement situations as well as to large inference problems. This is verified using synthetic data with linear and nonlinear signal models and in a nonlinear astronomical measurement setup.

This publication is structured as follows: In Section 2, we assume the setting of the generalized Wiener filter: a linear measurement equation, for a Gaussian signal sensed under Gaussian noise.<sup>[14]</sup> There, the optimal compression for prior mean reduces to an eigenvalue problem. In Section 3, this is generalized to nonlinear and non-Gaussian measurements. Furthermore, we show how a sparse structure can be utilized on the compression algorithm making it possible to handle large datasets in a reasonable amount of time. In Section 4, BDC is applied to synthetic data resulting from a linear measurement in one dimension, a nonlinear measurement in two dimensions, and real data from the Giant Metrewave Radio Telescope (GMRT).

## 2. Linear Compression Algorithm

We approach the problem of compression from a probabilistic perspective. To this end, we juxtapose the posterior probability distribution of the full inference problem with a posterior coming from a virtual likelihood together with the same prior. The goal is to derive an algorithm which takes the original likelihood and the prior as input and returns a new, virtual likelihood that is computationally less expensive than the original likelihood. This shall happen such that the resulting posterior probability distribution differs as little as possible from the original posterior.

The natural measure to compare the information content of a probability distribution and an approximation to it in the absence of other clearly defined loss functions is the KL divergence, as shown in ref. [12]. Minimizing the KL divergence completely leads to the criteria for the most informative likelihood.

### 2.1. Assumptions and General Problem

The new likelihood needs to be parameterized such that the KL divergence can be minimized. Initially, we make the following assumptions:

1. The signal  $s$ , which is a priori Gaussian distributed with known covariance  $S$ , has been measured with a linear response function  $R_o$ . The resulting original data  $d_o$  is subject to additive Gaussian noise with known covariance  $N_o$ . In summary,

$$d_o := R_o s + n_o \quad (1)$$

where we denote definitions by “:=”, with “:” standing at the side of the new defined variable, and  $s \leftrightarrow \mathcal{G}(s, S)$  and  $n_o \leftrightarrow \mathcal{G}(n_o, N_o)$  are drawn from zero centered Gaussian distributions. The notation  $s \leftrightarrow \mathcal{G}(s - s_o, S)$  indicates that  $s$  is drawn from a Gaussian distribution with mean  $s_o$  and covariance  $S$ . For the signal prior,  $s_o$  is zero.

2. The compressed data  $d_c$ , which is going to be lower dimensional compared to the original data  $d_o$ , is related to the signal

$s$  linearly through a measurement process with additive Gaussian noise with covariance  $N_c$  and response  $R_c$ , which need to be determined,

$$d_c = R_c s + n_c \quad (2)$$

In this setup, the likelihoods  $\mathcal{P}(d_i|s)$  as well as the posteriors of both the original and the compressed inference problem  $\mathcal{P}_i(s) := \mathcal{P}(s|d_i)$  are Gaussian again<sup>[14]</sup> (once  $N_c$  and  $R_c$  are specified):

$$\mathcal{P}(d_i|s) = \mathcal{G}(d_i - R_i s, N_i) \quad (3)$$

$$\mathcal{P}_i(s) = \mathcal{G}(s - m_i, D_i), \quad i \in \{o, c\} \quad (4)$$

The mean  $m_i$  and covariance  $D_i$  are

$$m_i := D_i R_i^\dagger N_i^{-1} d_i \quad (5)$$

and

$$D_i := (S^{-1} + M_i)^{-1}, \quad \text{with} \quad (6)$$

$$M_i := R_i^\dagger N_i^{-1} R_i \quad (7)$$

We call  $M_i$  the measurement precision matrix. Our goal is to find the compressed measurement parameters  $(d_c, R_c, N_c)$  such that the least amount of information on the signal  $s$  is lost as compared to  $(d_o, R_o, N_o)$ .

This means we want to adjust  $d_c$ ,  $R_c$ , and  $N_c$ , such that the difference of the two posteriors of the signal, given the compressed data  $d_c$  and given the original data  $d_o$ , is minimal. To this end, we minimize the KL divergence under the constraint that the compressed data vector shall not exceed a certain number of dimensions  $k_c$ :

$$\begin{aligned} \text{KL}_{o,c} &:= D_{\text{KL}}(\mathcal{P}_o || \mathcal{P}_c) \\ &:= \int ds \mathcal{P}(s|d_o) \ln \frac{\mathcal{P}(s|d_o)}{\mathcal{P}(s|d_c)} \\ &:= \langle \ln \mathcal{P}_o \rangle_{\mathcal{P}_o} - \langle \ln \mathcal{P}_c \rangle_{\mathcal{P}_o} \end{aligned} \quad (8)$$

In this notation,  $k_c$  is suppressed. It will become explicit in the next section. A detailed derivation showing that the KL divergence is indeed the appropriate measure to decide on the optimality of the compression and a discussion about the order of its arguments can be found in ref. [12].

For Gaussian posteriors, the KL divergence becomes

$$\begin{aligned} \text{KL}_{o,c}(d_c, R_c, N_c) &= \frac{1}{2} \text{tr} [D_c^{-1} D_o - \mathbb{1} - \ln(D_c^{-1} D_o) \\ &\quad + D_c^{-1} (m_c - m_o)(m_c - m_o)^\dagger] \end{aligned} \quad (9)$$

where the compressed posterior mean  $m_c$  and covariance  $D_c$  depend on the compressed measurement parameters  $d_c$ ,  $R_c$ , and  $N_c$  through (5) and (6). The superscript  $\dagger$  denotes the adjoint, that is, the transposed complex conjugate of a vector or linear operator.

Thus, we have formulated the original compression problem as a minimization problem:

$$d_c, R_c, N_c = \operatorname{argmin}_{d_c, R_c, N_c} \operatorname{KL}_{o,c}(d_c, R_c, N_c) \quad (10)$$

In order to arrive at the optimal choice of compressed measurement parameters, we minimize  $\operatorname{KL}_{o,c}$  sequentially with respect to its arguments  $d_c$ ,  $R_c$ , and  $N_c$ . In that procedure, we keep not yet optimized parameters as given and express already optimized parameters as functions of their given parameters during their optimization. Minimization of (9) with respect to the compressed data  $d_c$  for given response  $R_c$  and compressed noise covariance  $N_c$  yields:

$$\begin{aligned} d_c(R_c, N_c) &= N_c(R_c D_c R_c^\dagger)^{-1} R_c m_o \\ &= (R_c S R_c^\dagger + N_c)(R_c S R_c^\dagger)^{-1} R_c m_o \end{aligned} \quad (11)$$

using the identity

$$D R^\dagger = S R^\dagger (R S R^\dagger + N)^{-1} N. \quad (12)$$

Defining the compressed and original Wiener filter operator

$$\begin{aligned} W_i &:= S R_i^\dagger (R_i S R_i^\dagger + N_i)^{-1} \\ &= D_i R_i^\dagger N_i^{-1}, \quad i \in \{o, c\} \end{aligned} \quad (13)$$

we see that (11) is equivalent to  $R_c W_c d_c = R_c W_o d_o$  or  $R_c m_c = R_c m_o$ , meaning that the original and compressed posterior means are indistinguishable for the compressed response.

Inserting Equation (11) back into (9), we can define the already  $d_c$ -minimized  $\operatorname{KL}_{o,c}$ , which only depends on the still to be optimized parameters  $R_c$  and  $N_c$ :

$$\begin{aligned} \operatorname{KL}_{o,c}(R_c, N_c) &:= \operatorname{KL}_{o,c}(d_c(R_c, N_c), R_c, N_c) \\ &\hat{=} \frac{1}{2} (\operatorname{tr}[M_c D_o - \ln(S^{-1} + M_c)] \\ &\quad - m_o^\dagger R_c^\dagger (R_c S R_c^\dagger)^{-1} R_c m_o) \end{aligned} \quad (14)$$

where equalities up to constants that are irrelevant for the minimization, since they do not change the position of the minimum of  $\operatorname{KL}_{o,c}$ , are denoted by “ $\hat{=}$ ”. However, note that the specific values of  $\operatorname{KL}_{o,c}$  do not equal the divergence of the two posteriors anymore.

Only the last term of this new  $\operatorname{KL}_{o,c}$  in Equation (14) depends on the original data. The more the compressed response  $R_c$  preserves the information of the original posterior mean  $m_o$ , the smaller this term becomes which reduces  $\operatorname{KL}_{o,c}$ . Thus, a response  $R_c$  that is sensitive to  $m_o$  is favored. The original posterior mean  $m_o$  will typically exhibit large absolute values in signal space where the original response was of largest absolute values. This gives an incentive for the compressed response toward the original one.

In this section, we looked at the posterior distributions in the context of Gaussian prior and Gaussian likelihood in a linear measurement setup. By minimizing the Kullback–Leibler divergence with respect to the compressed data, we found an

expression for the latter. Plugging in this expression,  $\operatorname{KL}_{o,c}$  only depends on the compressed response and noise covariance.

## 2.2. Information Gain from Compressed Data

The remaining task is to minimize  $\operatorname{KL}_{o,c}(R_c, N_c)$  with respect to the compressed response  $R_c$  and noise  $N_c$ . Both appear in the loss function of our choice only combined in the measurement precision matrix  $M_c = R_c^\dagger N_c^{-1} R_c$  and as  $R_c$  only in the last term. Let  $U$  be a unitary transformation that diagonalizes the compressed noise covariance  $N_c$ . One can show that the transformation

$$R_c \rightarrow R'_c = U R_c \quad (15)$$

$$N_c \rightarrow N'_c = U N_c U^\dagger$$

leaves (14) invariant:

$$\begin{aligned} m_o^\dagger R_c^\dagger (R'_c S R_c^\dagger)^{-1} R'_c m_o &= m_o^\dagger R_c^\dagger U^\dagger U (R_c S R_c^\dagger)^{-1} U^\dagger U R_c m_o \\ &= m_o^\dagger R_c^\dagger (R_c S R_c^\dagger)^{-1} R_c m_o \end{aligned} \quad (16)$$

and analogous for the terms containing  $M'_c = R_c^\dagger U^\dagger U N_c^{-1} U^\dagger U R_c = R_c^\dagger N_c^{-1} R_c = M_c$ . We use this to find a parametrization of the compressed measurement precision matrix  $M_c$  with a set of vectors  $r := (r_i)_{i=1}^{k_c}$ , such that

$$M_c = \sum_{i=1}^{k_c} r_i r_i^\dagger \quad (17)$$

with  $k_c$  being the number of entries of the compressed data vector  $d_c$ . In its eigenbasis, the inverse compressed noise covariance reads

$$N_c^{-1} = \sum_{i=1}^{k_c} \mu_i^2 e_i e_i^\dagger \quad (18)$$

with  $e_i$  the normalized eigenvectors of  $N_c^{-1}$  and  $\mu_i^2$  the corresponding positive eigenvalues. For reasons that become clear later on, let us choose

$$R_c := \sum_{i=1}^{k_c} e_i \hat{r}_i^\dagger \quad (19)$$

$$\hat{r}_i := \frac{r_i}{\|r_i\|_S} \quad (20)$$

$$\mu_i = \|r_i\|_S \quad (21)$$

where  $\|\cdot\|_S$  is the norm induced by the prior covariance:  $\|x\|_S := \sqrt{x^\dagger S x}$ . With these definitions we can interpret  $\hat{r}_i$  as the (normalized) compressed measurement direction, that is, the direction in which the signal  $s$  is measured leading to the  $i$ th compressed data point.  $\hat{r} := (\hat{r}_i)_{i=1}^{k_c}$  shall then be orthonormal with respect to the scalar product induced by  $S$ . The noise of this measurement is given by the corresponding compressed noise

variance  $\mu_i$ . With these definitions (17) can be verified:

$$\begin{aligned} M_c &= R_c^\dagger N_c^{-1} R_c \\ &= \sum_{i,j,k=1}^{k_c} \hat{r}_i \underbrace{e_i^\dagger e_j}_{\delta_{ij}} \underbrace{e_j^\dagger e_k}_{\delta_{jk}} \hat{r}_k \mu_j^2 \\ &= \sum_{i=1}^{k_c} r_i r_i^\dagger \end{aligned} \quad (22)$$

Thus, the relevant degrees of freedom of the compressed response  $R_c$  and noise covariance  $N_c$  are encoded in the compressed measurement directions  $r = (r_i)_{i=1}^{k_c}$  and we can write

$$\begin{aligned} \text{KL}_{o,c}(r) &:= \text{KL}_{o,c}(R_c(r), N_c(r)) \\ &\hat{=} \frac{1}{2} \left( \text{tr} \left[ \left( \sum_i r_i r_i^\dagger \right) D_o - \ln \left( S^{-1} + \sum_i r_i r_i^\dagger \right) \right] \right. \\ &\quad \left. - m_o^\dagger R_c^\dagger (R_c S R_c^\dagger)^{-1} R_c m_o \right) \end{aligned} \quad (23)$$

We are left to evaluate the trace in (14). The main issue is the logarithmic term. By decomposing  $S = \sqrt{S} \sqrt{S}^\dagger$  it can be simplified:

$$\text{tr} \ln \left( S^{-1} + \sum_i r_i r_i^\dagger \right) \hat{=} \text{tr} \ln \left( \mathbb{1} + \sum_i \sqrt{S}^\dagger r_i r_i^\dagger \sqrt{S} \right) \quad (24)$$

We call  $\sqrt{S}$  prior dispersion. This nomenclature deviates from the common convention of using the names dispersion and covariance synonymously. An excitation  $\xi$  following a standard normal distribution  $\mathcal{G}(s, \mathbb{1})$  can be amplified by  $\sqrt{S}$ , that is,  $\sqrt{S}\xi$ . This amplified field follows the same statistics as  $s$ . We thus distinguish between variables defined in signal space and excitation space, with the dispersion  $\sqrt{S}$  (resp.  $\sqrt{S}^\dagger$ ) and its inverse serving as transformations between both.

The compressed measurement directions in excitation space are then

$$\begin{aligned} w_i &:= \sqrt{S}^\dagger r_i \\ \hat{w}_i &:= \frac{w_i}{\|w_i\|}, \quad \forall i \in \{1, \dots, k_c\} \end{aligned} \quad (25)$$

The orthonormal basis  $\hat{w} := (\hat{w}_i)_{i=1}^{k_c}$  diagonalizes both summands of (24) simultaneously:

$$\begin{aligned} \text{tr} \ln \left( \mathbb{1} + \sum_i \sqrt{S}^\dagger r_i r_i^\dagger \sqrt{S} \right) &= \text{tr} \ln \left( \mathbb{1} + \sum_i \|w_i\|^2 \hat{w}_i \hat{w}_i^\dagger \right) \\ &= \text{tr} \left[ \sum_i \ln \left( 1 + \|w_i\|^2 \right) \hat{w}_i \hat{w}_i^\dagger \right] \\ &= \sum_i \ln \left( 1 + \|w_i\|^2 \right) \\ &= \sum_i \ln \left( 1 + r_i^\dagger S r_i \right) \end{aligned} \quad (26)$$

In addition, the last term of (23) reduces to

$$\begin{aligned} m_o^\dagger R_c^\dagger (R_c S R_c^\dagger)^{-1} R_c m_o &= m_o^\dagger \sum_i \hat{r}_i e_i^\dagger \left( \sum_{j,k} e_j \underbrace{r_j^\dagger S r_k}_{\delta_{jk}} e_k^\dagger \right)^{-1} \sum_l e_l \hat{r}_l^\dagger m_o \\ &= m_o^\dagger \sum_{i,l} \hat{r}_i \underbrace{e_i^\dagger e_l}_{\delta_{il}} \hat{r}_l^\dagger m_o \\ &= m_o^\dagger \sum_i \hat{r}_i \hat{r}_i^\dagger m_o \end{aligned} \quad (27)$$

such that we get

$$\begin{aligned} \text{KL}_{o,c}(r) &\hat{=} \frac{1}{2} \left( \text{tr} [S^{-1} m_o m_o^\dagger] + \sum_{i=1}^{k_c} [r_i^\dagger D_o r_i - \ln(1 + r_i^\dagger S r_i) - m_o^\dagger \hat{r}_i \hat{r}_i^\dagger m_o] \right) \\ &\hat{=} - \sum_{i=1}^{k_c} \Delta I(r_i) \end{aligned} \quad (28)$$

with information gain  $\Delta I(r_i)$  of a single compressed measurement direction  $r_i$ . Since the compressed measurement directions in excitation space are orthogonal with respect to the standard scalar product, we proceed with the calculations in excitation space. There, the information gain becomes

$$\Delta I(w_i) := \frac{1}{2} [-w_i^\dagger D_o w_i + \ln(1 + w_i^\dagger w_i) + \hat{w}_i^\dagger \tilde{m}_o \tilde{m}_o^\dagger \hat{w}_i] \quad (29)$$

with posterior mean and covariance in excitation space

$$\tilde{m}_o := \sqrt{S}^{-1} m_o \quad (30)$$

$$D_o := \sqrt{S}^{-1} D_o \sqrt{S}^\dagger \quad (31)$$

$$\sqrt{S}^{-\dagger} := \sqrt{S}^{\dagger^{-1}}$$

We see that the relevant part of  $\text{KL}_{o,c}$  splits up into a sum over independent contributions  $-\Delta I(w_i)$  associated to the individual compressed measurement direction  $w_i$ , each of which belongs to a specific compressed data point  $(d_c)_i$ . Since  $\text{KL}_{o,c}$  expressed this way is additive with respect to the inclusion of additional data points, the sum in (28) can easily be extended. To minimize  $\text{KL}_{o,c}$  with respect to  $w$  (or  $r$ , respectively), the contributions  $(-\Delta I(w_i))$  can be minimized individually with respect to their respective compressed measurement direction  $w_i$ .

The information gain  $\Delta I$  depends on the normalized measurement direction  $\hat{w}_n$  and its magnitude  $\|w_n\|$ ,

$$\Delta I(\hat{w}_n, \|w_n\|) := \frac{1}{2} [-\|w_n\|^2 \hat{w}_n^\dagger D_o \hat{w}_n + \ln(1 + \|w_n\|^2) + \hat{w}_n^\dagger \tilde{m}_o \tilde{m}_o^\dagger \hat{w}_n] \quad (32)$$

We maximize this with respect to the magnitude, get

$$\|w_n\|^2 = \frac{1}{\hat{w}_n^\dagger D_o \hat{w}_n} - 1 \quad (33)$$

and insert the result into (32). This leaves us with our final expression of the information gain of a single compressed data point:

$$2\Delta I(\hat{w}_n) = \hat{w}_n^\dagger D_o \hat{w}_n - 1 - \ln(\hat{w}_n^\dagger D_o \hat{w}_n) + \hat{w}_n^\dagger \tilde{m}_o \tilde{m}_o^\dagger \hat{w}_n. \quad (34)$$

Summarizing, the sets of vectors  $\{e_i\}$  and  $\{r_i\}$  are representations of the compressed noise covariance and compressed response, such that the trace in  $KL_{o,c}$  splits up into independent summands. Each summand is the negative information gain when considering the corresponding compressed measurement direction  $\hat{w}_n$  and can be treated individually. As a next step, we need to maximize  $\Delta I(\hat{w}_n)$ .

### 2.3. Optimal Expected Information Gain

In order to find the compressed data point which adds most information to the compressed likelihood, (34) needs to be maximized with respect to the normalized vector  $\hat{w}_n$ . For zero posterior mean  $m_o = 0$ , this problem reduces to an eigenvalue problem as shown in Appendix A. We proceed by treating the general case of non-vanishing  $m_o$ .

There is only one normalized vector  $\hat{w}_i$  (respectively  $\hat{r}_i$ ) left to be determined for each  $i \in \{1, \dots, k_c\}$ . However, in the current form, (34) cannot be maximized analytically. The main issue is the last term that contains  $\tilde{m}_o$  and which in the following we treat stochastically using the prior signal and noise knowledge on the signal  $s$  and the noise  $n_o$  only. Thereby, we can calculate the expected information gain.

Using the Gaussian priors of signal  $s$  and noise  $n_o$  with zero mean, as well as the measurement (1) and the definition (5) of  $m_o$ , we get the expected posterior signal mean

$$\langle m_o \rangle_{P(s, n_o)} = D_o R_o^\dagger N_o^{-1} (R_o \underbrace{\langle s \rangle_{P(s)}}_{=0} + \underbrace{\langle n_o \rangle_{P(n_o)}}_{=0}) = 0 \quad (35)$$

and variance

$$C := \langle m_o m_o^\dagger \rangle_{P(s, n_o)} \quad (36)$$

$$\begin{aligned} &= D_o R_o^\dagger N_o^{-1} (R_o S R_o^\dagger + N_o) N_o^{-1} R_o D_o \\ &= (D_o R_o^\dagger N_o^{-1} (R_o S R_o^\dagger + N_o) N_o^{-1} R_o + \mathbb{1}) D_o - D_o \end{aligned}$$

$$\stackrel{(12)}{=} (S R_o^\dagger N_o^{-1} R_o + \mathbb{1}) D_o - D_o$$

$$\stackrel{(6)}{=} S - D_o \quad (37)$$

Thus, as a sum of Gaussian distributed variables,  $m_o$  again is Gaussian distributed with

$$P(m_o) = \mathcal{G}(m_o, C) \quad (38)$$

Calculating the mean of the last term of (34) under this distribution then gives

$$\begin{aligned} \hat{w}_n^\dagger \langle \tilde{m}_o \tilde{m}_o^\dagger \rangle_{P(m_o)} \hat{w}_n &= \hat{w}_n^\dagger \sqrt{S^{-1}} C \sqrt{S^{-1}} \hat{w}_n \\ &\stackrel{(37)}{=} \hat{w}_n^\dagger (\mathbb{1} - D_o) \hat{w}_n \end{aligned}$$

$$\stackrel{(25)}{=} 1 - \hat{w}_n^\dagger D_o \hat{w}_n \quad (39)$$

which cancels the first two terms. The expected information gain then is

$$\langle \Delta I(\hat{w}_n) \rangle_{P(m_o)} = -\frac{1}{2} \ln(\hat{w}_n^\dagger D_o \hat{w}_n) \quad (40)$$

This expected information gain is maximal, if and only if  $\hat{w}_n$  is parallel to the eigenvector of  $D_o$  with smallest eigenvalue  $\delta_n^2$ . This insight reduces the problem to the eigenvalue problem

$$D_o w_n = \delta_n^2 w_n. \quad (41)$$

In terms of the vectors  $r$ , which build the compressed measurement precision matrix  $M_c$  in (17), and after inserting (6), this states

$$\begin{aligned} D_o^{-1} S r_n &= (\mathbb{1} + M_o S) r_n \\ &= \delta_n^{-2} r_n \end{aligned} \quad (42)$$

Combining (21), (33), and (41) gives:

$$\mu_n^2 = \|w_n\|^2 = \delta_n^{-2} - 1 \quad (43)$$

Thus (42) is equivalent to

$$M_o S r_n = \mu_n^2 r_n \quad (44)$$

We call  $M_o S =: Q$  the *fidelity matrix*. For identity responses,  $Q$  can be interpreted as signal-to-noise covariance ratio. The largest eigenvalues of  $Q$  give rise to the most informative compressed measurement directions according to the minimization of the Kullback–Leibler divergence. At the same time,  $Q$  is the matrix product of the original measurement precision matrix  $M_o$  and the signal prior covariance  $S$ . Thus, its largest eigenvalues and corresponding eigenvectors are those directions in signal space, where the measurement is maximally precise while the prior is maximally uncertain. In other words, the directions at which the original data update the prior the most, are exactly those, which are the most informative.

For  $r_n = R_o^\dagger v$ , eigenvalue problem (44) is equivalent to the generalized eigenvalue problem of ref. [13],

$$R_o S R_o^\dagger v = \lambda N_o v \quad (45)$$

multiplied with  $R_o^\dagger N_o^{-1}$  from the left.

For constant noise  $N_o = \sigma_{n_o}^2 \mathbb{1}$ , eigenvalue problem (44) becomes a Bayesian analog to principal component analysis (PCA).<sup>[9,10]</sup> PCA takes the highest eigenvalues and corresponding eigenvectors of the data covariance matrix. In PCA, this data covariance is built by the covariance of several measurements. In a Bayesian setting, we can determine the data covariance with prior information using Equation (1) instead:

$$\langle d_o d_o^\dagger \rangle_{P(s, n)} = R_o S R_o + N_o \quad (46)$$

with corresponding eigenvalue problem

$$(R_o S R_o^\dagger + N_o) \nu = \lambda \nu \quad (47)$$

Multiplying with  $N_o^{-1}$  from the left gives;

$$(N_o^{-1} R_o S R_o^\dagger + \mathbb{1}) \nu = \lambda N_o^{-1} \nu \quad (48)$$

We can subtract  $\nu$  on both sides of the equation and for constant noise  $\sigma_n^2$  identify  $\mu_n^2$  with  $\lambda \sigma_n^{-2} - 1$ , such that

$$N_o^{-1} R_o S R_o^\dagger \nu = \underbrace{(\lambda \sigma_n^{-2} - 1)}_{\mu_n^2} \nu \quad (49)$$

where we inserted the constant noise value only on the right hand side of the equation to illustrate the similarity to Equation (44) on the right hand side. Multiplying with  $R_o^\dagger$  from the left and identifying  $r_n = R_o^\dagger \nu$  this gives Equation (44). In contrast to original PCA, in a Bayesian setting, we can already compress a single measurement. In contrast to this Bayesian analog to PCA, BDC is able to handle varying noise and compresses optimally with respect to the information about the variable of interest  $s$  as the result of our derivation.

With Equation (44), the expected information gain for including the compressed data point  $d_{c_n}$  in BDC then is

$$\begin{aligned} \Delta I(\mu_n) &:= \langle \Delta I(\hat{r}_n) \rangle_{P(m_o)} \\ &= -\frac{1}{2} \ln(\hat{w}_n \underbrace{D_o \hat{w}_n}_{=\delta_n^2 \hat{w}_n}) \\ &= -\ln(\delta_n) \\ &= \frac{1}{2} \ln(\mu_n^2 + 1) \end{aligned} \quad (50)$$

Summarizing, we need to find the  $k_c$  largest eigenvalues  $\mu_n^2$  and corresponding eigenvectors  $r_n$  of (44). With these, the compressed measurement parameters are

$$N_c^{-1} = \sum_{n=1}^{k_c} \mu_n^2 e_n e_n^\dagger \quad (51)$$

$$R_c = \sum_{n=1}^{k_c} e_n \hat{r}_n^\dagger \quad (52)$$

$$d_c = (R_c S R_c^\dagger + N_c) (R_c S R_c^\dagger)^{-1} R_c m_o \quad (53)$$

with

$$\hat{r}_n := \frac{r_n}{\|r_n\|_S} \quad \text{and} \quad \|r_n\|_S = \mu_n \quad (54)$$

These equations and (5) and (6) are all ingredients needed to solve the compression problem.

## 2.4. Algorithm

Now, the previously derived method shall be turned into the actual BDC. For that we need to solve eigenvalue problem (44).

For compressing the data to  $k_c$  data points, one needs to determine the  $k_c$  largest eigenvalues and corresponding eigenvectors that belong to the most informative measurement directions. First we derive an estimate for the fraction of information stored in the compressed measurement parameters if we compute only a limited number of eigenpairs, that is, eigenvalues and corresponding eigenvectors. Then, we discuss some details of how to compute the input parameters for getting the compressed measurement parameters, that is, for the eigenvalue problem (44) and how to solve it.

Due to computational limits, in general we cannot determine all  $K$  eigenpairs of (44) carrying information. We need to set the number  $k_{\max}$  of most informative eigenpairs being determined numerically. For that limited number of eigenpairs, we derive a lower bound of the information stored in the corresponding compressed measurement parameters in the following. If we are only interested in a certain amount of information we can use this bound to find and neglect eigenpairs containing too little information, such that in the end, we have  $k_c \leq k_{\max}$  eigenpairs containing still enough information.

The eigenpairs carrying information are those with non-zero eigenvalue  $\mu_i$ . The number  $K$  of non-zero eigenvectors is equal to the rank of  $M_o S = R_o^\dagger N_o^{-1} R_o S$ . As Gaussian covariances,  $S$  and  $N_o$  are positive definite and therefore have full rank, the original response  $R_o$  has at most a rank equal to the smaller rank of both covariances. Thus, with (7), the rank of  $M_o S$  and therefore the number of informative eigenpairs  $K$  is equal to the rank of  $R_o$ . Altogether the compressed measurement parameters can maximally carry the total information  $I := \sum_{i=1}^K \Delta I(\mu_i)$ .  $I$  is the difference in information stored in the posterior, when considering all informative eigendirections  $k_c = K$  compared to having no compressed data,  $k_c = 0$ . For no compressed data, the compressed posterior distribution becomes the prior distribution with covariance  $S$ . Thus,  $I$  is the total information about the signal  $s$  encoded in the original data with respect to prior knowledge.

We can find an upper bound for the total information. The eigenpairs are ordered such that the eigenvalues decrease with growing index, and therefore the contribution to the total information sum  $I$ . Thus, the last determined eigenvalue is the smallest of all eigenvalues and the least informative one. Also, it is larger than all eigenvalues that could not be computed. We can use this to give an upper limit bound to the amount of information lost by truncating the sum at  $k_c$ , and adding the number of eigenvalues ignored,  $K - k_{\max}$ , times the amount of information provided by the last eigenpair  $I(\mu_{k_{\max}})$ . Thus,

$$I \leq \sum_{i=1}^{k_{\max}} \Delta I(\mu_i) + (K - k_{\max}) \Delta I(\mu_{k_{\max}}) \quad (55)$$

We define the fraction of information  $\gamma$  of total information  $I$  stored in the compressed measurement parameters which are determined by the eigenpairs:

$$\gamma := \frac{\sum_{i=1}^{k_c} \Delta I(\mu_i)}{I} \quad (56)$$

With Equation (55), we can find a lower bound

$$\begin{aligned} \gamma &\stackrel{(55)}{\geq} \frac{\sum_{i=1}^{k_c} \Delta I(\mu_i)}{\sum_{i=1}^{k_{\max}} \Delta I(\mu_i) + (K - k_{\max}) \Delta I(\mu_{k_{\max}})} \\ &=: \gamma_{\min}(k_c, (\mu_i)_{i=1}^{k_{\max}}, K) \end{aligned} \quad (57)$$

Using  $I \geq \sum_{i=1}^{k_{\max}} \Delta I(\mu_i)$ , we analogously find an upper bound

$$\gamma \leq \gamma_{\max} := \frac{\sum_{i=1}^{k_c} \Delta I(\mu_i)}{\sum_{i=1}^{k_{\max}} \Delta I(\mu_i)} \quad (58)$$

for  $\gamma$ . Those bounds can be used to narrow the fraction of information stored in the compressed measurement parameters  $\gamma$  by  $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ .

Now we can find the minimum number of eigenpairs containing at least  $\gamma_{\min} I$  information by finding the smallest number of eigenpairs  $k_c$  such that (57) holds and then forget all eigenpairs with a larger index than  $k_c$ .

In case  $k_{\max} > K$ , some eigenpairs contain no additional information to the prior knowledge and the information gain of the last eigenpair  $\Delta I(\mu_{k_{\max}})$  is zero. Then we have stored all information ( $\gamma = 1$ ) in the compressed data with non-zero eigenvalue and Equation (57) is automatically fulfilled for given  $\gamma_{\min}$ .

With the information fraction  $\gamma$ , we have found a quantification of how much of the available information is stored in the compressed measurement parameters. For limited number of computed eigenpairs, one can still estimate  $\gamma$  by its upper and lower bounds. Next, we discuss some details of the computation of the input parameters of BDC and its final implementation.

When evaluating (5) for the original posterior mean  $m_o$  we have to avoid the inversion of the prior covariance  $S$  or the prior dispersion  $\sqrt{S}$  as the explicit inversion of a  $n \times n$  matrix is of  $\mathcal{O}(n^3)$ . Such expensive operations can be partly avoided by getting  $m_o$  via

$$\begin{aligned} m_o &= \sqrt{S} \tilde{m}_o \\ &= \sqrt{S} D_o \tilde{j}_o \\ &= \sqrt{S} (\mathbb{1} + \sqrt{S}^\dagger R_o^\dagger N_o^{-1} R_o \sqrt{S})^{-1} \sqrt{S}^\dagger R_o^\dagger N_o^{-1} d_o \end{aligned} \quad (59)$$

Compared to directly calculating  $m_o$  with Equations (5) and (6), this saves one inversion of the prior covariance  $S^{-1}$ . Still the linear operator  $(\mathbb{1} + \sqrt{S}^\dagger R_o^\dagger N_o^{-1} R_o \sqrt{S})$  needs to be inverted. In this case, we can make use of the conjugate gradient algorithm which computes the application of the inverse of a matrix to a vector in  $\mathcal{O}(n)$ .

The eigenvalue problem (44) can then be solved by an Arnoldi iteration.<sup>[15]</sup>

This leaves us with basic BDC as summarized in Algorithm 1. We first compute the original posterior mean and prior covariance using the prior dispersion. Given the original data  $d_o$ , response  $R_o$  and inverse noise covariance  $N_o^{-1}$ , we can compute the fidelity matrix to solve eigenvalue problem (44) for the  $k_{\max}$  largest eigenvalues. If a minimal amount of information fraction that shall be encoded in the compressed data is specified, we can determine the largest index  $k_c$  so that Equation (57) holds and

#### Algorithm 1 Basic Bayesian Data Compression

---

```

1:   procedure compress  $\sqrt{S}$ ,  $R_o$ ,  $N_o^{-1}$ ,  $d_o$ ,  $k_{\max}$ ,  $\gamma_{\min}$ 
2:      $m_o = \sqrt{S} (\mathbb{1} + \sqrt{S}^\dagger R_o^\dagger N_o^{-1} R_o \sqrt{S})^{-1} \sqrt{S}^\dagger R_o^\dagger N_o^{-1} d_o$ 
3:      $S = \sqrt{S} \sqrt{S}^\dagger$ 
4:     compute largest eigenpairs  $(\mu_i^2, r_i)_{i=1}^{k_{\max}}$  of  $R_o^\dagger N_o^{-1} R_o S$ 
5:     find smallest  $k_c$ , such that (57) holds.
6:     for every  $i > k_c$  do
7:       forget  $(\mu_i^2, r_i)$ 
8:     for every  $i \leq k_c$  do
9:        $r_i \leftarrow \frac{r_i}{\sqrt{r_i^\dagger S r_i}}$ 
10:     $R_c = \sum_{i=1}^{k_c} e_i r_i^\dagger$ , with unit vectors  $\{e_i\}_{i=1}^{k_c}$ 
11:     $N_c^{-1} = \text{diag}((\mu_i^2)_{i=1}^{k_c})$ 
12:     $d_c = (R_c S R_c^\dagger + N_c) (R_c S R_c^\dagger)^{-1} R_c m_o$ 
13:    return  $d_c$ ,  $R_c$ ,  $N_c^{-1}$ 

```

---

only save those  $k_c$  eigenpairs that carry that much information. Then we normalize the eigenvectors with respect to the norm induced by the prior covariance  $\|\cdot\|_S$  to finally determine the compressed measurement parameters  $(d_c, R_c, N_c)$  according to Equations (51)–(53).

In the linear scenario the full Wiener filter needs to be solved. Thus, the computational resources required to compute and store the compressed measurement parameters exceed the resources saved by the compression. It would be of benefit in a real world application if the eigenfunctions could be re-used in repetitions of the same measurement and do not need to be computed again. BDC's main benefit lies in the nonlinear scenario with a nonlinear response inside the measurement equation. There, the inference appears to be more complicated, but BDC enables us to exploit information stored in the data further while calling the original data and response less often.

## 3. Generalizations

### 3.1. Generalization to Nonlinear Case

The derivation of BDC so far is based on a linear measurement equation. In real world problems, however, often nonlinear measurement equations

$$d = R(s) + n \quad (60)$$

describe the relation of signal and data. There, the response transforms the signal nonlinearly. In addition, the signal parameters can be very non-Gaussian and inter-dependent through a deep hierarchical model. In those cases, we need to adjust basic BDC. Unlike basic BDC in the linear case, the adjusted BDC in nonlinear scenarios can save computation time as the original data and response do not need to be called as often as in the full reconstruction.

Let us assume that such complications are expressed via a deep hierarchical model. Following ref. [16], deep hierarchical models can be transformed into independent standard normal distributed parameters by encoding prior knowledge into the likelihood. In this fashion, the complexity of a deep hierarchical model

is stored in a nonlinear function  $f$ . This function relates the parameters  $s$  of the hierarchical model—the actual signal—to the parameters  $\xi$  of a transformed, flattened, non-hierarchical model via

$$s = f(\xi) \quad (61)$$

The transformation  $f$  has to be chosen such that the prior of  $\xi$  becomes a standard normal distribution:

$$\mathcal{P}(d|s)\mathcal{P}(s)ds = \mathcal{P}(d|f(\xi))\mathcal{G}(\xi, \mathbb{1})d\xi \quad (62)$$

Thus, we call  $\xi$  excitation field as defined in Section 2.2.

For flattened models we can now deal with nonlinear measurement setups using metric Gaussian variational inference (MGVI).<sup>[17]</sup> There, the posterior is approximated by a Gaussian  $\mathcal{G}(\xi - \bar{\xi}, \Xi)$  with inverse Fisher information metric as uncertainty covariance  $\Xi$  centered on some mean value  $\bar{\xi}$ . The Fisher information metric is

$$M_{d|s} = \left\langle \frac{\partial \mathcal{H}(d|s)}{\partial s^\dagger} \frac{\partial \mathcal{H}(d|s)}{\partial s} \right\rangle_{\mathcal{P}(d|s)} \quad (63)$$

Here

$$H(d|s) := -\ln \mathcal{P}(d|s) \quad (64)$$

is the information Hamiltonian of the likelihood. In order to distinguish the approximate uncertainty  $\Xi$  from the true posterior uncertainty covariance, we call it variational uncertainty. With a standard Gaussian prior, the posterior covariance then states

$$\Xi^{-1} = J_\xi^\dagger M_{d|s} J_\xi + \mathbb{1} \quad (65)$$

with Jacobian

$$J_\xi := \frac{\partial f(\xi)}{\partial \xi} \quad (66)$$

For many measurement situations, the response splits into a linear part  $R_{\text{lin}}$  and a nonlinear part  $R_{\text{nl}}$

$$d_o = R_{\text{lin}} R_{\text{nl}}(s) + n_o \quad (67)$$

The linear part might describe a linear telescope response, or just be an identity operator. Then, we redefine our signal  $s' := R_{\text{nl}}(s)$ . Before,  $s$  were the parameters of the hierarchical model with a function  $f$  transforming the standard Gaussian distributed excitation  $\xi$  into  $s$ . Now  $s'$  are our parameters being related to  $\xi$  via  $R_{\text{nl}} \circ f$ . Thus, the Jacobian states

$$J'_\xi = \frac{\partial R_{\text{nl}}(f(\xi))}{\partial \xi} \quad (68)$$

and we define

$$\sqrt{S'} := \left. \frac{\partial R_{\text{nl}}(f(\xi))}{\partial \xi} \right|_{\xi=\bar{\xi}} \quad (69)$$

This is a linearization of the nonlinear part evaluated at  $\bar{\xi}$ , a reference value of  $\xi$ , for example, the current mean location provided

by the MGVI algorithm, such that

$$J'^\dagger_\xi M_{d|s} J'_\xi = \sqrt{S'}^\dagger M_{d|s} \sqrt{S'} \quad (70)$$

The compression is then applied to the linear measurement equation

$$d_o = R_{\text{lin}} s' + n_o \quad (71)$$

with given noise covariance  $N_o$ . This way, we have all the ingredients for BDC to work in the nonlinear case as well.

During the inference process, the approximated mean  $\bar{\xi}$ , at which the linearization is evaluated, changes. With updated knowledge also the compression input will change. This suggests the following strategy:

1. Compress the original measurement parameters with prior knowledge and original measurement parameters as input.
2. Infer the posterior mean given the compressed measurement parameters. This will only be an approximate solution.
3. Approximate the original posterior around the inferred mean and use it as the new prior to start again with the first step.

We will call the number of compressions, that is, the number of total repetitions of those three steps,  $n_{\text{comp}}$ , the number of MGVI minimization steps to infer the mean with compressed data in between  $n_{\text{rep}}$ . In total the original data and response only have to be used as often as in  $n_{\text{comp}}$  minimization steps, while in total we reach  $n_{\text{comp}} \times n_{\text{rep}}$  minimization steps exploiting the information in the data. A crucial step will be to find the optimal exploration ( $n_{\text{comp}}$ ) versus exploitation ( $n_{\text{rep}}$ ) ratio as in ref. [18].

### 3.2. Utilization of Sparsity

High-dimensional data are difficult to handle simultaneously. For the eigenvalue problem of BDC, it is more efficient to solve a larger number of lower dimensional problems. For the signal inference, it is beneficial to ensure that the response  $R_c$  and noise  $N_c$  of the compressed system are sparse operators. This can be achieved by dividing the data into patches to be compressed separately. For that, we use the fact that not every data point carries information about all degrees of freedom of the signal at once. Data points that inform about the same degrees of freedom of the signal can then be compressed together exploiting sparsity of the compressed measurement directions. This also has the advantage of lower dimensional eigenvalue problems to be solved, saving computation time. The separately compressed data of the patches as well as corresponding responses and noise covariances are finally concatenated.

An example would be data and signal that are connected via a linear mask hiding parts of the signal from the data as discussed in Section 4.2. If then the signal is correlated in space, we can divide the data into patches which carry information about the same patch in signal space.

Alternatively, this method can be used to compress data online, that is, while data is measured one can collect and process it blockwise as suggested by ref. [19] such that the full data never

has to be stored completely. After each compression, the reconstruction of the signal takes the concatenated measurement parameters, where the compressed response is now sparse, and solves the inference problem altogether.

Mathematically speaking, we divide the original data  $d_o$  into separated sets of data  $d_{o_i}$  with responses  $R_{o_i}$  and noise covariance  $N_{o_i}$  for every patch  $i$ . The responses  $R_{o_i}$  are already sparse not covering the whole signal space as such. Then we compress those datasets separately leading to  $d_{c_i}$ ,  $R_{c_i}$ , and  $N_{c_i}$ . Concatenating them back again leads to the final measurement equation

$$\underbrace{\begin{pmatrix} d_{c1} \\ \vdots \\ d_{cn} \end{pmatrix}}_{d_c} = \underbrace{\begin{pmatrix} R_{c1} \\ \vdots \\ R_{cn} \end{pmatrix}}_{R_c} s + \underbrace{\begin{pmatrix} n_{c1} \\ \vdots \\ n_{cn} \end{pmatrix}}_{n_c} \quad (72)$$

with noise covariance

$$N_c = \begin{pmatrix} N_{c1} & & \\ & \ddots & \\ & & N_{cn} \end{pmatrix} \quad (73)$$

We call this process patchwise compression. If the compression of all original data is done at once, we call it joint compression. With patchwise compression, signal correlations between data points of different patches cannot be exploited for the compression. One should aim to assign strongly correlated data to the same patches such that their correlation is considered in BDC. For data correlated in space, correlations are strongest between data for neighboring signal locations. It makes sense to choose the patches by vicinity. The computational benefit due to sparsity of the response contrasts the information loss due to patching. One can increase the dimension of the compressed data to compensate that loss and still requires less storage capacity.

The signal is not affected by the patching. Signal correlations are still represented via the signal prior covariance  $S$ , and therefore also present in the compressed signal posterior. Since the reconstruction is running over the full problem, its result is not biased due to patchwise compression. In principle, any kind of compression could be specified via introduction of arbitrary  $R_c$  and  $N_c$  into Equation (11). The resulting reconstruction would all be unbiased, but of course, less accurate.

To summarize, we separate the data into patches. The data of every patch is compressed separately leading to compressed measurement parameters for every patch. Prerequisite for treating the patches separately is that the noise of the individual patches is uncorrelated between the patches. By concatenating the compressed measurement parameters of all patches, we get all operators needed for the compressed signal posterior. This removes the need to store the compressed responses over the entire signal domains. Only their patch values have to be stored, saving memory and computation time.

## 4. Application

Now, the performance of BDC is discussed for applications of increasing complexity, first for a linear synthetic measurement setting and then for a nonlinear one. For the latter, we

demonstrate the advantage of dividing the data into patches and compressing them separately. Finally, the compression of radio interferometric data from the GMRT is discussed.

### 4.1. Synthetic Data: Linear Case

First, the BDC is applied to synthetic data in the Wiener filter context. This means all probability distributions such as prior, likelihood, and posterior are Gaussian and the data are connected to the signal via a linear measurement equation  $d = Rs + n$ . In this setup, we can test basic BDC in its actual, not approximated form, for changing noise and masked areas. Also, we compare it with the Bayesian analog of PCA (BaPCA), reducing the expected data covariance to its principal components.

The signal domain is a 1D regular grid with 256 pixels. The synthetic signal and corresponding synthetic data are drawn from a zero centered Gaussian prior. The data is masked, such that only pixels 35–45 and 60–90 are measured linearly, according to  $d = Rs + n$ . Additionally, white Gaussian noise is added with zero mean and standard deviation of  $\sigma_n = 2 \times 10^{-3}$  for measurements up to pixel 79, and  $4 \times 10^{-3}$  for pixels 80–90. Those noisy data are then compressed to four data points, from which the signal is inferred in a last step.

The signal covariance is assumed to be diagonal in Fourier space, with the power spectrum

$$P_s(k) := \frac{2 \cdot 10^4}{1 + \left(\frac{k}{20}\right)^4} \quad (74)$$

The signal itself can be computed from the power spectrum via

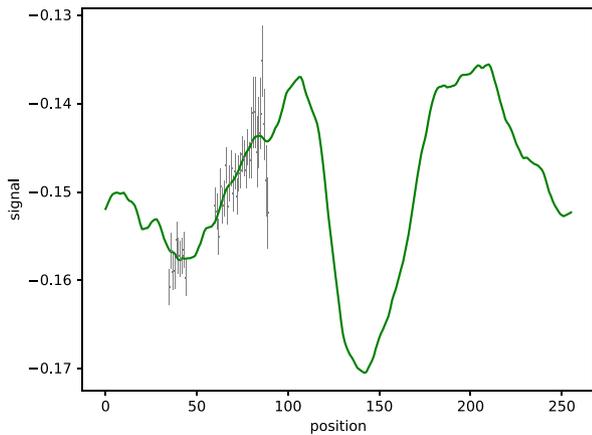
$$s = F \sqrt{P_s(k)} \xi_k \quad (75)$$

with a Fourier or Hartley transformation  $F$  and the Fourier modes  $\xi_k$  being drawn independently from a standard Gaussian  $\mathcal{G}(\xi, 1)$ . The response is set to be a mask measuring pixels 35–45 and 60–90 directly, with a local and thereby unity response. The measurement setup with signal mean, synthetic signal, and data are shown in **Figure 1**.

We apply basic BDC as described in Algorithm 1. For the eigenvalue problem, we use the implementation of the Arnoldi method in `scipy` (`scipy.sparse.linalg.eigs`<sup>[20]</sup>).

After having compressed the data, we evaluate the reconstruction performance using the compressed data. With Equation (4) the posterior can be calculated directly from the compressed measurement parameters, and signal covariance. The posterior mean and uncertainty for the original and the compressed data are compared to the ground truth in **Figure 2**. The original data has been compressed from 40 to 4 data points with a fraction  $\gamma$  of 83.7% of the total information encoded in the compressed data. Especially at the measured areas, both the original and the compressed reconstruction are close to the ground truth, while the reconstructed means deviate from the ground truth at masked areas far away from measured areas. However, this deviation is still captured in the uncertainties.

Figures 2 and 3 show that the compressed posterior has a higher variance than the original posterior. **Figure 3** shows



**Figure 1.** 1D synthetic data setup to test BDC. The synthetic signal is marked in green and the measured data in gray. Error bars at the data points mark the Gaussian noise standard deviation of  $2 \times 10^{-3}$  for pixels before 79 and  $4 \times 10^{-3}$  for pixels 80–90. Those noisy data shall be compressed by BDC and used for reconstructing the synthetic signal back again.

the relative uncertainty difference of the compressed and original posterior, that is, the compressed posterior uncertainty subtracted from the original posterior uncertainty divided by their mean at each pixel. It is strictly positive in the measured areas as well as in the unmeasured areas. In the measured area compared to the unmeasured area, the relative uncertainty excess of the compressed posterior is higher, since there the absolute uncertainty is low. Slight absolute increase of uncertainty there leads to a higher relative variation. This proves the increase of uncertainty due to the compression.

The eigenvectors are plotted in Figure 3. At the masked pixels, the eigenvectors stay zero. The changing noise covariance visibly impacts the shape of the eigenvectors. Between pixels 79 and 80, where the noise increases, is a clear break in all the eigenvectors. A higher noise standard deviation leads to abrupt drops in the eigenvectors. A more detailed discussion of the eigenvectors can be found in Appendix B. There we compare the shape of the eigenvectors in a simpler setup of a continuous mask and constant noise to those of Chebyshev polynomials of the first kind. An analytical derivation of their form in this simple setting is given in Appendix C.

For comparison, we compressed the data of this setup with BaPCA defined by eigenvalue problem (47) and performed a linear Wiener filter. Note that in the original PCA, there is no noise defined. We determined the largest four eigenvalues from Equation (47) and corresponding eigenvectors  $v$ . Those eigenvectors were used as the row vectors of a transformation  $V$  which compresses the original data. The measurement equation for those compressed data then becomes

$$\begin{aligned} d_{\text{pca}} &= Vd_o \\ &= VR_o s + Vn_o \end{aligned} \quad (76)$$

Identifying  $R_{\text{pca}} := VR_o$  as the response of this compressed system and  $n_{\text{pca}} := Vn_o$  as its noise, one can apply a linear Wiener filter as in Equations (5)–(7).

We have plotted the corresponding reconstructions in Figure 2 with corresponding uncertainty as cyan dotted line with horizontally hatched shades. BaPCA reconstructs the mean and standard deviation similar to BDC. For comparison, we have also plotted the relative uncertainty in Figure 3 as we did for BDC as well as the eigenvectors building the compressing transformation  $V$ . The relative uncertainty from BaPCA clearly exceeds the one of BDC in areas of low noise. In the area of higher noise, the relative uncertainty excess from BaPCA is lower than the one from BDC compared to the original posterior uncertainty. The reason for this can be seen by comparing the eigenvectors of BaPCA and BDC: BaPCA is more sensitive in high noise areas, therefore having a lower posterior uncertainty there, but also letting thereby more noise enter the compressed data. Compared to BaPCA, BDC encodes more information in regions of lower noise, where the data is more informative, and it keeps less information from regions of higher noise.

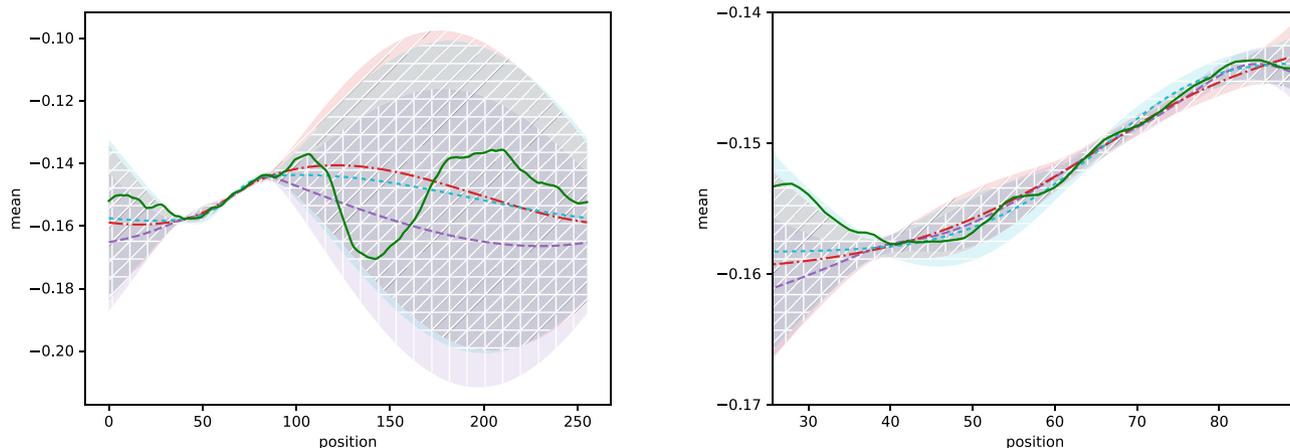
This can be seen when looking at the eigenvectors of both methods. The amplitude of eigenvectors from BDC drop where the noise standard deviation becomes higher. For BaPCA, only the fourth eigenvector changes its amplitude in the region of higher noise standard deviation. The first three eigenvectors, however, do not vary their amplitude with noise change. This coincides with the observation that BaPCA and BDC become equivalent for constant noise standard deviation as discussed in Section 2.3.

We found that the compression method reduces the dimension of the data with minimal loss of information in the simple case of a linear 1D Wiener filter inference. Storing only four compressed data points still reconstructs the signal well, compared to the reconstruction with original data. Every compressed data point determines the amplitude of an eigenvector such that the signal is approximated appropriately. The lossy compression leads to a slightly higher uncertainty, as information is lost. In this application, BDC and BaPCA give similar results in terms of reconstruction. Compared to a BaPCA, BDC focusses more on regions of lower noise standard deviation, where the data are more informative.

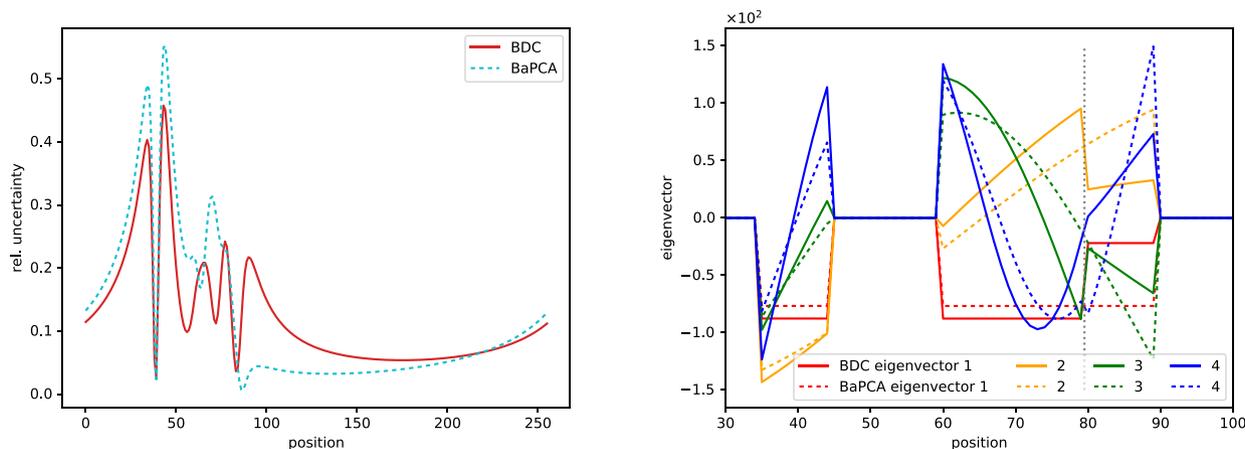
## 4.2. Synthetic Data: Nonlinear Case

Testing BDC on data from a nonlinear generated signal in two dimensions allows us to verify the derivation of the nonlinear approximation in Section 3.1 and also test patchwise compression discussed in Section 3.2. Some nonlinear synthetic signal is generated and then inferred with the original data, with compressed data, and with data, which has been first divided into patches and then compressed. The results are discussed with respect to the quality of the inferred mean for the different methods, their standard deviation, their power spectrum, as well as the computation time.

The synthetic signal has been generated with a power spectrum created by a nonlinear amplitude model as described in ref. [21] deformed by a sigmoid function to create a nonlinear relation between signal and data. The code of the implementation can be found here: <https://gitlab.mpcdf.mpg.de/jharthki/bdc>. The resulting ground truth lies on an  $128 \times 128$  regular grid and is shown on the top left most panel of Figure 4. To test



**Figure 2.** Inference of synthetic signal (green) in a linear Wiener filter setup. The posterior mean given the original data shown in Figure 1 is plotted as dashed purple line with vertically hatched shades marking its standard deviation. The same for the posterior mean given four compressed data points from BDC as dash-dotted red lines with diagonally hatched shades and for the posterior mean given four compressed data points from BaPCA as dotted cyan lines with horizontally hatched shades. The left graph shows the full domain and the right one a zoom in the measured area. The left and right subfigures show the full domain and the region constrained by data, respectively. For all methods, the ground truth is consistent with the  $1\sigma$  uncertainty.

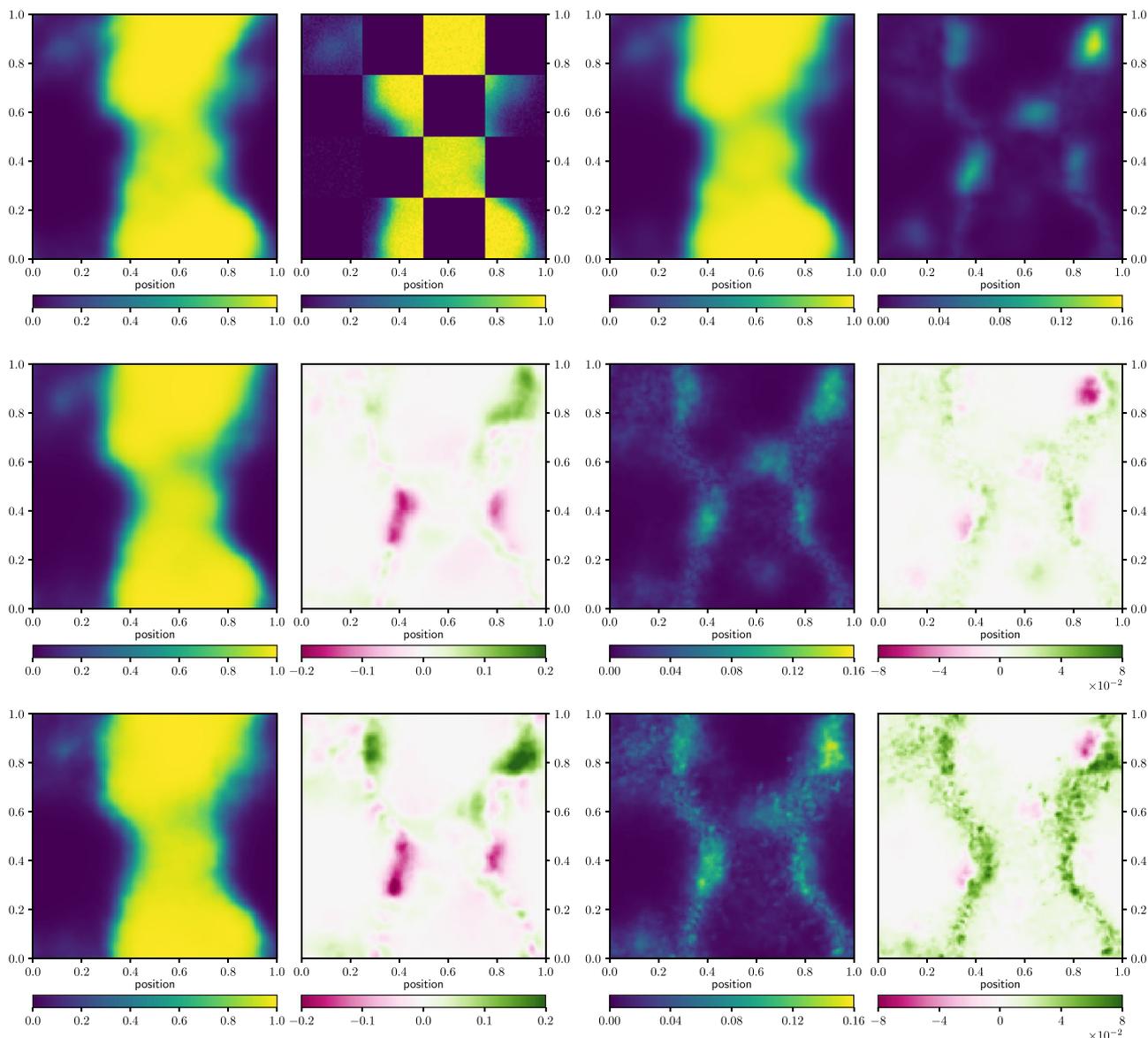


**Figure 3.** Left: Relative uncertainty excess of the compressed reconstruction with respect to the one based on the original data in red. The same for the reconstruction from BaPCA compressed data with a dotted cyan line. Due to information loss the uncertainty of the BDC reconstruction is higher than the original one leading to positive values only. In the measured area, the relative uncertainty excess rises. The BaPCA reconstructed uncertainty has a similar slope with a higher uncertainty excess in regions with smaller noise standard deviation and smaller uncertainty excess in regions with higher noise standard deviation. Right: Relevant part of the eigenvectors of the eigenvalue problem (44) (BDC, solid lines) and of eigenvalue problem (47) (BaPCA, dashed lines) for the synthetic data setup in Figure 1 over the signal domain. The eigenvectors are colored in red, orange, green, and blue in descending order of their eigenvalues. The gray dotted line marks the noise level change. All eigenvectors are zero in the unmeasured area. In addition the ones from BDC decline in the area of higher noise while only the fourth eigenvector of BaPCA slightly changes its slope depending on the noise. A detailed discussion of the shape of BDC eigenvectors can be found in Appendix B.

BDC on masked areas in a nonlinear setup as well, this signal is covered by a  $4 \times 4$  checkerboard mask with equally sized  $32 \times 32$  squares, as displayed in the second top panel of Figure 4. Additionally uncorrelated noise with zero mean and 0.02 standard deviation has been added. From this incomplete and noisy data, the non-Gaussian signal as well as the power spectrum of the underlying Gaussian process need to be inferred simultaneously. The results of the original inference are plotted in the third top panel of Figure 4.

After setting up the input parameters for BDC, the data was compressed from 8 192 to  $k_{\max} = 80$  data points altogether without sorting out less informative data points. Next metric Gaussian

variational inference<sup>[17]</sup> performed  $n_{\text{rep}} = 2$  inference steps based on the compressed data, each time finding a better approximation for the posterior mean and approximating the posterior distribution again. Then the original data was compressed another time using the current posterior mean as the reference point  $\bar{\xi}$ . This was done  $n_{\text{comp}} = 3$  times in total. To determine the amount of information contained in the resulting compressed data another run has been started where 4096 eigenpairs have been computed. This way the estimation of  $\gamma$  is more exact using Equation (57) for a lower bound  $\gamma_{\min}$  and (58) for an upper bound  $\gamma_{\max}$ . With this we estimate that the compressed data of size 80 contains 31.4–32.2% of the information. The same way, one gets that 672 compressed

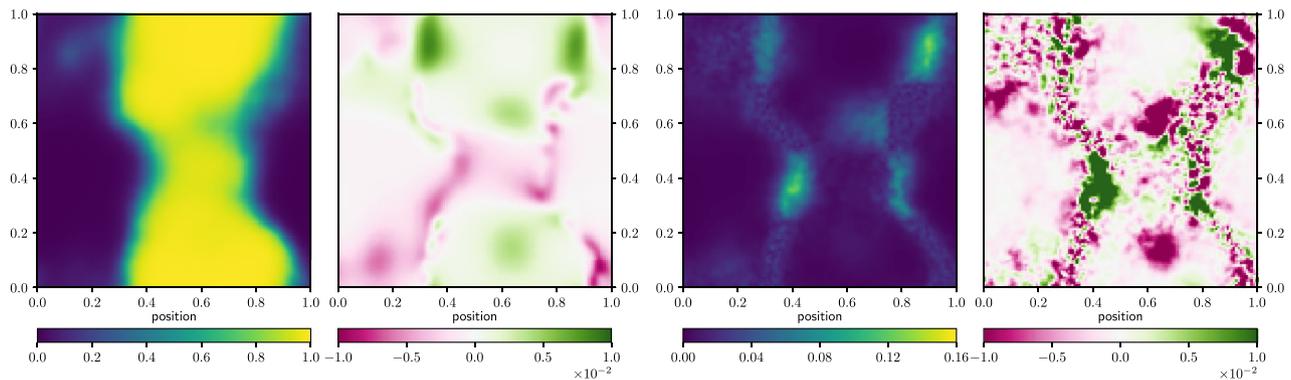


**Figure 4.** Top row: From left to right: First, the generated ground truth and second, the synthetic data for a synthetic signal created by the amplitude model and processed by a sigmoid function. A unity response connects the data and the signal with a four times four checkerboard mask hiding parts of the signal from the data. Third, using MGVI, the posterior mean of the signal and fourth, its uncertainty as inferred with the original data. This setup is used to test BDC. The reconstruction from original data serves as a reference for the reconstruction from compressed data. Middle row: Results for jointly compressed data. Three compression and MGVI steps were performed. First, the posterior mean. Second, its difference to the mean of the inference without compression, that is, the compressed posterior mean subtracted from the original posterior mean. Third, the posterior variance of the inference with compressed data. And fourth, the same, subtracted from the variance of the posterior with the original data. The reconstructions from original and jointly compressed data deviate mainly at the edges of larger structures, at most at the order of  $10^{-1}$ . The compressed posterior uncertainty is higher than the one of the original posterior at most locations. Bottom row: As in the middle, just with patchwise compressed data. As in the case of reconstruction from jointly compressed data, the posterior mean mainly deviates at the edges of large structures of the signal from the original posterior mean. Again, the standard deviation is mainly higher for the compressed posterior compared to the original one.

data points contain 80% of the information. It turns out that already 80 compressed data points contain enough information to reconstruct the essential structures of the signal as one can see from reconstruction and difference maps shown in Figure 4.

The corresponding posterior mean is plotted in the center left of Figure 4 together with the difference to the originally inferred mean. Overall the compression yields similar results. Deviations

appear at the edges of homogeneous structures, while deviations inside homogeneous structures are neglectable. The variance is plotted in the center right of Figure 4. Again it differs mainly at the edges. Since during the compression process information is lost, the results should have higher uncertainty in general. This is almost everywhere the case, however, there are some parts, which report a better significance than the reconstruction



**Figure 5.** From left to right: Mean computed from the compressed measurement parameters directly from Equation (5) without another minimization using MGVI. Second, its difference to the mean computed from the compressed data by MGVI as in Figure 4. Third and forth, the same for the uncertainty computed from Equation (6). The structures are maintained well. Deviations are at the order of  $10^{-2}$ . This is one magnitude lower than the standard deviation.

without compression. This either implies an inaccuracy of BDC or that BDC can partly compensate the approximation MGVI brings into the inference by providing it with measurement parameters that are better formatted for its operation.

Let us discuss in more detail how such a high loss in information still reproduces reasonable results. The information loss is equivalent to a widening of the posterior distribution. Its quantitative value in terms of  $(1 - \gamma)$  does not take into account on which scales the information is mainly lost. As can be observed, a substantial fraction of the measurement information constrains small scales. Losing this part does not make such a large difference to the human eye, in particular as the small scale structures are of smaller amplitude, but the information loss is measured on relative changes. Thus a loss of 70% of mainly small scale information is possible without increasing the error budget significantly.

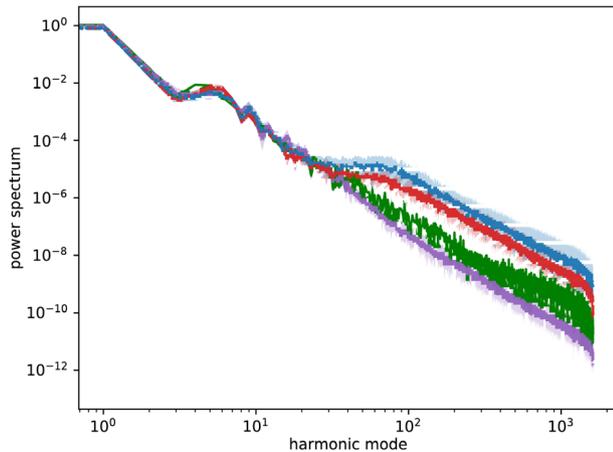
Now we investigate patchwise compression in the same setup. The data in each of the eight measured squares were compressed to  $k_{\max} = 10$  data points separately. In total there are as many compressed data points for the patchwise compression as for the joint one. We can use that the response only masks the signal but does not transform it. Thus, we can compress the data with prior information of the corresponding patch only. This way we reduce the dimension of the eigenvalue problem (44) to the size of the patch. Before the reconstruction, the resulting compressed responses are expanded to the full signal space in a sparse form and concatenated as described in Section 3.2. For the reconstruction, the whole signal is inferred altogether. The resulting mean and variance of the inference for this method and their difference to the original ones are shown in the lower part of Figure 4. Both differ mainly at the edges of homogeneous structures from the original posterior mean and variance. Also for the case of patchwise compressed data, the variance at some points becomes smaller than for the original inference. Since patchwise BDC does not use knowledge about correlations between the patches for the compression, it compresses less optimal than joint BDC, and thus mostly has a higher deviation of the mean and higher uncertainty in the reconstruction. Like in the case of compressed data, we improve the estimation of  $\gamma$  by computing 4096 eigenpairs, that is, 512 eigenpairs per patch. It turns out that for every patch

on average 34.8–35.4% of information is kept about the signal inside this patch when using the 10 most informative eigenpairs for the compressed data points, where all patches but two contain 15–25%. Counting from left to right and top to bottom, the data compressed from the second and fifth patch with very homogeneous signal contain more than 75% of the information. When comparing the  $\gamma$  values of each patch, one needs to consider that the patches are compressed individually. Therefore the information of one and another patch might be partly redundant and their individual  $\gamma$ s can not just be added or averaged in order to get the joint information content.

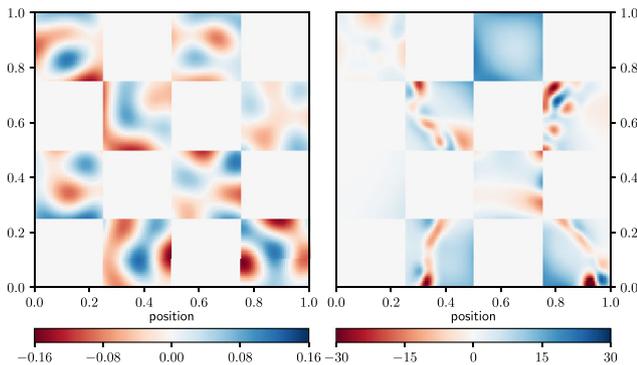
After having computed the compressed measurement parameters, we can also directly get the compressed posterior mean and covariance from Equations (5) and (6). The inference from the compressed data then reduces to a linear Wiener filter problem. The resulting mean and variance are plotted in Figure 5 together with their difference to the mean and variance using one more MGVI inference. Both deviate at the order of  $10^{-2}$  which is one magnitude lower than the uncertainty. This illustrates that the compression helped to linearize the inference problem around the posterior mean.

Finally the results of the methods can be compared by looking at the inferred power spectra of the underlying Gaussian process in Figure 6. All of the reconstructions recover the power spectrum well for high harmonic modes up to the order of  $10^1$ . For higher modes, the samples of the originally inferred power spectrum tend to lie below the ground truth. In contrast, the reconstructions of the two compression methods overestimate the power spectrum for higher modes. It is not completely clear why this is the case. For higher modes, the signal to noise ratio is low. In those regimes it is more difficult to reconstruct the power spectra. This could be a reason for the deviation on high harmonic scales. In addition variational inference methods tend to underestimate uncertainties.<sup>[22]</sup> Since we use MGVI for the reconstruction for all methods, this could cause the inferred power spectra not to coincide within their uncertainties.

It is interesting to have a closer look on the back projection of the compressed data, that is,  $R_c^\dagger d_c$  as well as on the projection of the eigenvectors building the compressed responses onto the space. The back projection of the jointly compressed data before



**Figure 6.** The power spectrum of the underlying Gaussian process: ground truth (solid, green) and inferred from original (purple), jointly compressed (red) and patchwise compressed (blue) data. Qualitatively all power spectra reconstruct the ground truth well for small harmonic modes while they deviate at higher modes. There, the power spectrum reconstructed from original data tends to underestimate the amplitude of the original power spectrum while the power spectra reconstructed from jointly compressed data overestimate it. Compared to the joint compressed power spectrum, the patchwise compressed power spectrum deviates even more for higher harmonic modes. This inaccuracy contrasts the computational benefit of patchwise compression due to utilizing sparsity.



**Figure 7.** The back projected jointly compressed data before (left) and after (right) the inference. Before, the structure is rather uniformly distributed while the back projection of the compressed data clearly focuses on the edges of the signal structures after the inference.

the first inference, that is, having looked at the original data only once, is shown in **Figure 7** on the left. In contrast to the back projection after the minimization process in the right plot, the data look quite uninformative, covering more or less uniformly the whole probed signal domain. After the inference, when the reference point around which the linearization is made has changed, the jointly compressed data addresses mainly regions of rapid changes in the signal. Especially the contours at the edges are saved in the jointly compressed data. This is even clearer visible in the projection of the eigenvectors  $r_i$  building the compressed response according to (19) in **Figure 8**. The first two eigenvectors capture the frame of the large structure. The third one mainly looks at the upper left corner, where also some structure occurs, though it is less distinct than the large one. None of the eigenvectors

covers any structure in the second and fifth patches, which was also the patches with largest  $\gamma$ , that is, least information loss due to the compression. Since the structure of the ground truth there is rather uniform, it does not contain much information but the amplitude of the field, which can easily be compressed to few data points.

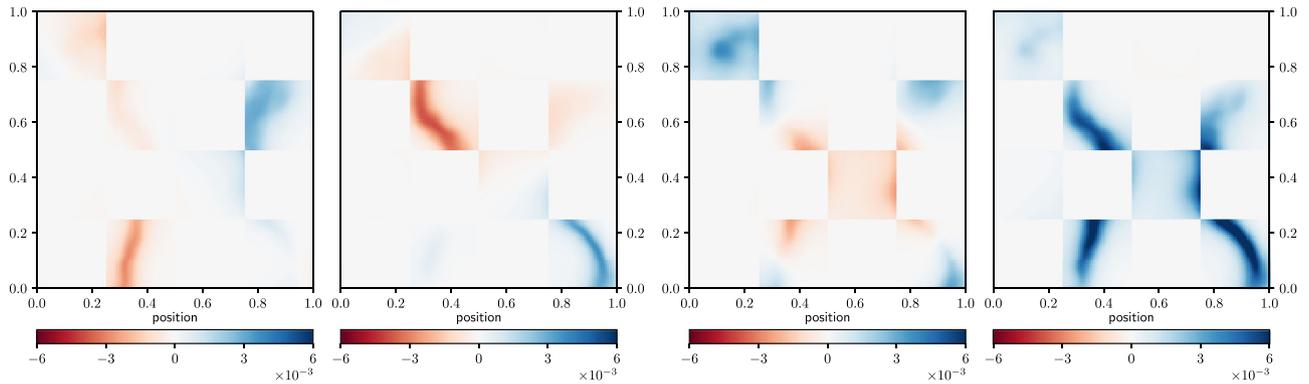
**Figure 9** shows the back projection of the patchwise compressed data before and after the inference. Here, the change of the basis functions becomes apparent as well.

**Table 1** shows the computation time of the different compression methods and reconstructions. As described above, it has been measured for  $(n_{\text{rep}}, n_{\text{comp}}) = (2, 3)$ , as well as  $(n_{\text{rep}}, n_{\text{comp}}) = (3, 2)$ . In case of the original inference, in total  $n_{\text{rep}} \times n_{\text{comp}}$  inference steps were performed such that in there is an identical number of inference steps for every method. The time has been measured for the inference only and for the total run of separation of the data into patches,  $n_{\text{comp}}$  compressions with  $n_{\text{rep}}$  inferences after each compression. The average of all  $n_{\text{rep}} \times n_{\text{comp}}$  inferences is given in the first line of **Table 1**. The time for the total runs is given in the second line. It has been measured on a single node of the FREYA computation facility of the Max Planck Computing & Data Facility restricted to 42 GB RAM. In all categories, the inference with patchwise compressed data is the fastest. In contrast, joint compression takes the longest time. There, one can clearly see the advantage of patching as discussed in **Section 3.2**. This leads to sparse responses which are more affordable in terms of computation time and storage. This is in agreement with the synthetic example discussed in this section. It shows that such sparse representations are highly beneficial.

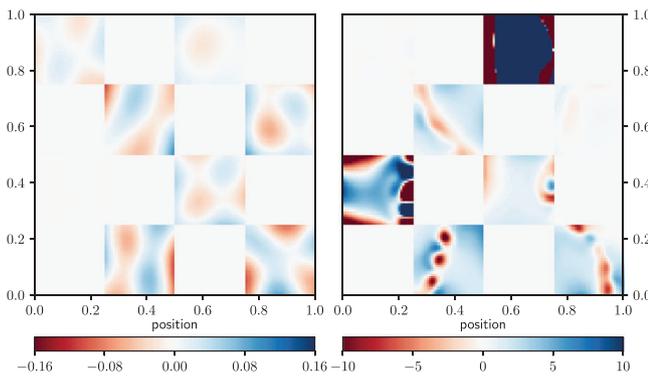
The response is called, that is, applied, several times during the minimization. In the application here, calling the response is inexpensive. However, there are applications in which the response is expensive. Then one aims to minimize the number of response calls, as this determines the computation time. The number of response calls were counted for  $(n_{\text{rep}}, n_{\text{comp}}) = (2, 3)$ . In the inference with the original data,  $R_0$  was called 686 187 times. In the process of compressing jointly it was called 4 369 times. During the patchwise compression, the patchwise original response has been called 10 776 times. In the case of patchwise compression, the response only maps between the single patches, that is, it is a factor 16 smaller than the full response. Thus, effectively the full original response has been called only about 674 times in the case of patchwise compression leading to a speed up factor of up to 1018 in case the response calculation is the dominant term. In this application, patchwise compression lead to computation times consistent with the computation time of inferring with original data. Future steps to make BDC more rewarding could be to find representations for the compressed response that are even more feasible.

### 4.3. Real Data: Radio Interferometry

Finally, we apply BDC to radio astronomical data from the supernova remnant Cassiopeia A observed by the GMRT.<sup>[23,24]</sup> 200 000 data points from the measurement were selected randomly and noise corrected according to [25]. Using those data, two images were constructed by the RESOLVE algorithm<sup>[25,26]</sup> that relies on MGVI, where we make one image with compression and one without compression for comparison.



**Figure 8.** The eigenvectors, that is, the row vectors of the jointly compressed response after the inference corresponding to the highest eigenvalues in descending order from left to right. They clearly capture the shape of the signal with higher amplitude at the edges of larger structures.



**Figure 9.** Back projection of the patchwise compressed data before (left) and after (right) the inference. As in the case of joint compression, structures in the signal can be recognized in the back projection of the compressed data after the inference. In addition the back projection is high in rather uniform areas of the signal.

**Table 1.** Computation times for the inference with original, compressed, and patchwise compressed data in the nonlinear synthetic application.

$(n_{\text{rep}}, n_{\text{comp}}) = (3, 2)$	Original	Comp	Patchcomp
Inference time [s]	317	984	289
Total run time [s]	633	1972	586
No response calls	686187	2913	6062
$(n_{\text{rep}}, n_{\text{comp}}) = (2, 3)$			
Inference time [s]	205	637	200
Total run time [s]	615	1917	613
No response calls	686187	4369	10776

Also the number of original response calls is stated. The original data has been inferred  $n_{\text{rep}} \times n_{\text{comp}}$  times.

The model used by RESOLVE as described in ref. [26] is adopted for the inference. To this end, the variables from the amplitude model in ref. [26] are denoted as  $\xi$  and transformed to the sky  $s$ —the actual signal—by  $s = f(\xi)$  as described in ref. [16]. The data are connected to the sky  $s$  via a nonlinear measurement equation of the form (67). The nonlinear part of the response  $R_{\text{nl}}$  contains a pointwise exponentiation and a Fourier transformation onto a grid, which leads to the variable

of interest for the compression method  $s'$ . The linear part of the response  $R_{\text{lin}}$  degrids the resulting points and transforms them to the data  $d_o$ , that is, it projects the points lying on a grid to continuous space. The total response  $R = R_{\text{lin}} \circ R_{\text{nl}}$  directly maps from the sky  $s$  to the data  $d_o$ , such that we have the measurement equations

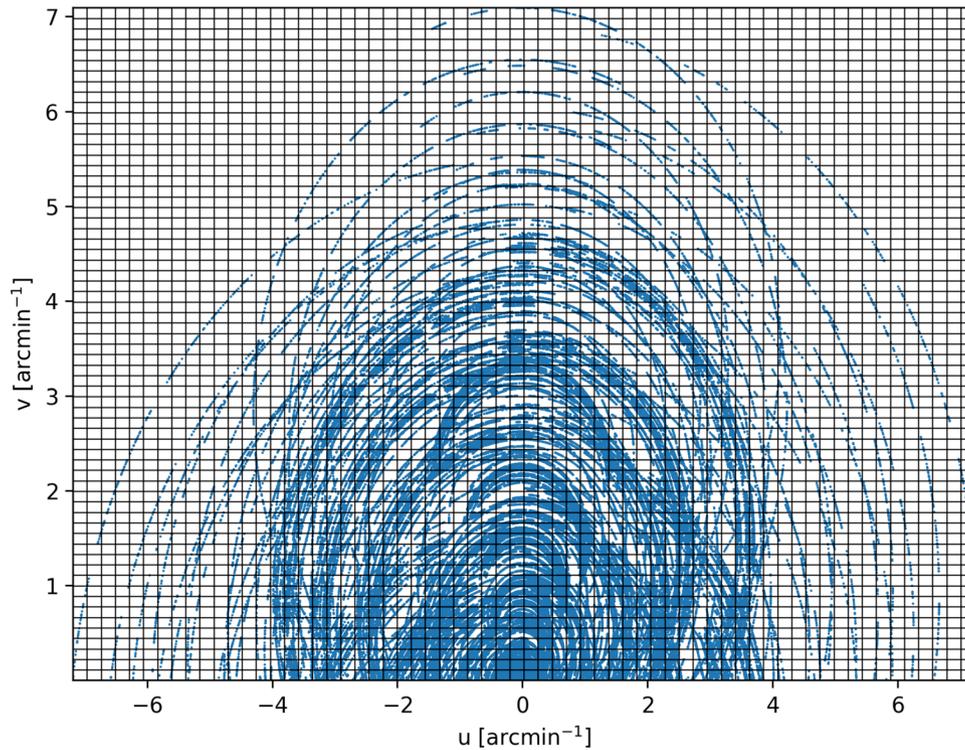
$$\begin{aligned} d_o &= R(s) + n \\ &= R_{\text{lin}}(R_{\text{nl}}(s)) + n \\ &=: R_{\text{lin}}(s') + n \end{aligned} \quad (77)$$

$R_{\text{lin}}$  is computationally expensive and the data  $d_o$  are large. The aim is to compress the original data with the signal  $s'$  as the variable of interest. It turns out that the joint compression of all data is infeasible due to its large computational costs. So, only the patchwise compression is tested. Similar to the previous example, a non-Gaussian signal and the power spectrum of the underlying Gaussian process need to be estimated simultaneously from noisy and very incomplete data, only here the nonlinearity is an exponential function and the sparse data live in Fourier space.

We divide the Fourier plane into  $64 \times 64$  squared patches, as shown in **Figure 10**. The location of the measurements in the Fourier plane are marked as well. It is apparent that many patches are free of data, some contain some data, and the highest density of data points occurs around the origin of the Fourier plane.

**Figure 11** shows the image obtained from the original dataset and from the patchwise compressed dataset. The mean is obtained after three MGVI iteration steps from the original data. Using BDC, the data in each patch was compressed ideally under prior information.

The resulting data, noise covariance, and responses were concatenated and used for inference in MGVI with three inference steps. The received posterior distribution was used to compress the separated original data once more with updated knowledge, inferred with three minimization steps. Doing this one more time resulted in the reconstruction shown here. The bottom right plot in **Figure 11** shows that the uncertainty of the reconstruction from patchwise compressed data is mostly higher than the uncertainty from the original reconstruction. This is expected due to the information loss of the compression. The data points



**Figure 10.** The position of all data points in Fourier space. The overlaid grid marks the data patches according to a Cartesian grid. Those patches are used for the patchwise compression.

in every patch have been compressed to maximally  $k_{\max} = 64$  data points per patch. The minimal fraction of information stored in the compressed measurement parameters  $\gamma_{\min}$  has been set to 0.99. In total, this leads to 73 239 compressed data points, which is a reduction of the data size by a factor 2.73. Due to the large computation time, it is infeasible to determine more eigenpairs. Thus, estimates of the amount of information  $\gamma$  contained in the compressed data points are very vast. On average, with Equations (57) and (58), the estimated range of  $\gamma$  is 0.35–94.8% for every patch with a standard deviation of 0.39% and 17.3%, respectively. The dispersion of those values for different patches is high, caused by the varying distribution of data points per patch.

The corresponding power spectra of the underlying Gaussian processes are shown in **Figure 12**. Their slopes qualitatively agree, but partly deviate outside their uncertainties. As discussed for the synthetic application in Section 4.2, one needs to take into account that MGVI tends to underestimate the uncertainties. In general, the power spectrum inferred from the original data is more distinct in its slope than the one from compressed data in terms of deviations from a straightly falling power spectrum. The compressed spectrum is flatter especially in the higher harmonic regime. This could be caused by a lower signal to noise ratio in this regime. However, it is not completely clear, why BDC shows this behavior.

In addition to Euclidean gridded patches, we separated the data into equiradial and equiangular patches. This leads to a more even distribution of data points inside the patches, since this way patches become larger further outside, where there are less data points. However, for this patch pattern small structures in the

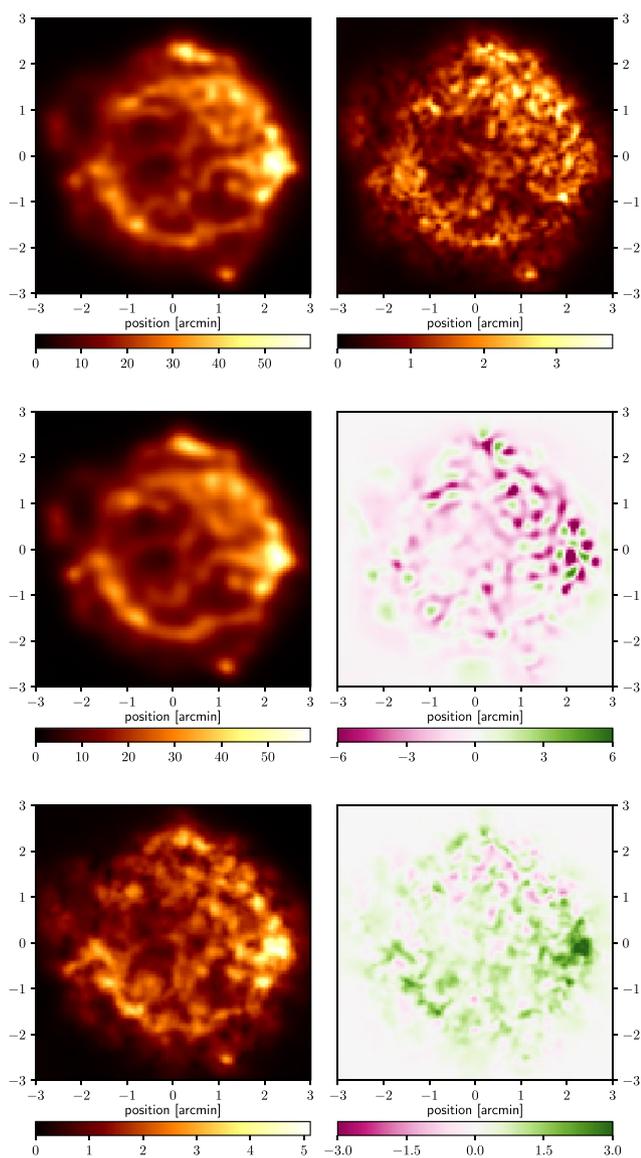
reconstruction get lost. The reason for this is that data points in the Fourier plane far away from the origin store information about the small scale image structures. Compressing them together therefore can be expected to lead to a loss of information on small scale structures. Thus, two criteria need to be considered for the choice of the patch geometry: from a computational perspective, patches with few data points are favored. From information theoretical perspective, data points carrying similar information shall be compressed together.

This application shows that BDC is able to operate on real world datasets in the framework of radioastronomical image reconstruction. The run time still can be improved, though. Since the compression for different patches works independently, this can perfectly be parallelized. Another potential area for improvement would be the choice of the separation of the patches. One needs to aim for a patch geometry, where the original data points are evenly distributed among the patches, while also highly correlated data points are assigned to the same patch.

## 5. Conclusion

A generic Bayesian data compression algorithm has been derived, which compresses data in an optimal way in the sense that as much information as possible about a signal for which the correlation structure is assumed to be known a-priori.

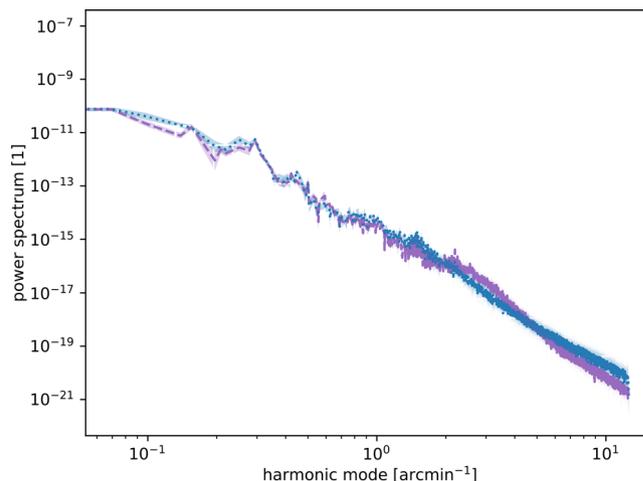
Our derivation is based on the Kullback–Leiber divergence. It reproduces the results of ref. [13] that optimizing the information loss function leads to a generalized eigenvalue problem. We generalized the method to the nonlinear case with the help of metric Gaussian variational inference.<sup>[17]</sup> Also, we divided the dataset



**Figure 11.** The reconstruction of supernova remnant Cassiopeia A from GMRT measurements. The colorbars have units  $\text{Jy arcmin}^{-2}$ . Top row: To the left the resulting posterior mean, to the right its posterior uncertainty using original data. Middle row: To the left reconstructed mean using patchwise compressed data. To the right its difference to the reconstructed mean from original data. The difference map shows that mainly small scale structures are lost due to the compression. Bottom row: The same for the posterior uncertainty. Since information is lost due to the compression, the uncertainty is mainly higher than the one of the original posterior.

into patches to limit the computational resources needed for the compression. This leads to sparseness of the response, allowing to apply the method in high-dimensional settings as well.

The method has been successfully applied to synthetic and real data problems. In an illustrative 1D synthetic linear scenario, 40 data points could be compressed to four data points with less than 20% loss of information. In a more complex, 2D and nonlinear synthetic measurement scenario, 8192 measurements could be reduced to 80 data points with 70% loss of information



**Figure 12.** Similar to Figure 6. The power spectra of the underlying Gaussian process inferred from original and patchwise compressed data are depicted in purple (dashed) and blue (dotted), respectively. They roughly agree in their slope.

that still capture the essential structures of the signal. Dividing the data into patches resulted in a huge reduction of the required computation time for the compression itself, confirming the expected advantage.

Finally, the method has been applied to real astrophysical data. The radio image of a supernova remnant has been reconstructed qualitatively with a data reduction by a factor of almost 3.

For such scientific applications of BDC, one needs to choose the variable of interest, the signal  $s$ , such that  $s$  represents the scientific interest best. Then BDC can optimally adjust which information need to be stored in the compressed data optimally. In the chosen example, all degrees of freedom of the field had to be stored. In principle, only certain (Fourier) scales or certain areas of a field could be defined to be the quantity of interest. This allows BDC to remove information on the other, irrelevant scales or areas.

BDC compresses optimally with respect to the knowledge about this quantity. It is a lossy compression method, that is, the compressed data contain less information about the quantity of interest—information that is relevant to answer the scientific question—than the original data. This information loss consistently leads to higher uncertainty in the linear case where the solution is exactly known and mostly higher uncertainty in the nonlinear applications where only approximate solutions can be found. To quantify the loss of information we have introduced the fraction  $\gamma$  of information about the quantity of interest  $s$  stored in the compressed data compared to the information in the full data. This fraction can be used to reduce the dimension of the compressed data such that they still contain the relevant amount of information. In case the compression is too lossy, one needs to adjust the number  $k_{\max}$  of computationally determined eigenpairs that build the compressed measurement parameters or increase the limit  $\gamma_{\min}$  of information that should be contained in the compressed data.

Still, the current BDC algorithm requires too much resources in terms of storage needed for the responses and computation time. In order to improve this even further, the choice of the

data patches can still be investigated and optimized such that data points storing similar information are assigned to the same patch. Up to now, data have been patched which are neighboring in real or Fourier space. However also data points could be informationally connected non-locally. One would need to look at the Kullback–Leibler divergence again to find those connections and group the data accordingly.

Another problem is the computational cost of the response. In the course of our derivation, we represented it in a vector decomposition. One could demand further restrictions to those vectors such as a certain parametrization or find other representations to find a computationally ideal basis for the responses. This could lead to a higher reduction factor in applications such as the astrophysical application in Section 4.3.

As a final step BDC needs to prove its advantage in real applications. A promising application could be online compression such as ref. [19] suggest. In a scenario, where data come in block-wise, those blocks can be treated as the patches and compressed separately. This is for example applicable in any experiment running over time. There, time periods imply the measurement blocks. This way the original data never needs to be stored at all, but compressed immediately and optimally under the current knowledge.

## Appendix A: Optimality of BDC for Zero Posterior Mean

In this section of the appendix, we prove that for zero original posterior mean  $m_o = 0$  the compression is optimal, if  $\hat{w}_i$  is the eigenvector to the smallest eigenvalue of  $D_o$ . Optimal means that the information gain  $2\Delta I(\hat{w}_i)$  in (34) is maximal with respect to  $\hat{w}_i$ .

*Proof.* For the proof, the  $\hat{w}_i$  dependence in (34) needs to be shown explicitly:

$$2\Delta I(\hat{w}_i) := \hat{w}_i^\dagger D_o \hat{w}_i - 1 - \ln(\hat{w}_i^\dagger D_o \hat{w}_i) \quad (\text{A1})$$

Let  $\{(v_i, \delta_i)\}$  be the eigenpairs of  $D_o$ . The eigenvectors  $\{v_i\}$  form a complete orthonormal basis. Then  $\hat{w}_i$  can be written as  $\hat{w}_i = \sum_j \omega_{ij} v_j$  with  $\sum_j \omega_{ij}^2 = 1$ , such that  $\hat{w}_i$  is normalized, and

$$\begin{aligned} 2\Delta I(\hat{w}_i) &= 2\Delta I\left(\sum_j \omega_{ij} v_j\right) \\ &= \sum_j \omega_{ij}^2 \delta_j - 1 - \ln\left(\sum_j \omega_{ij}^2 \delta_j\right) \end{aligned} \quad (\text{A2})$$

We will use that  $f(x) = x - 1 - \ln x$  is a convex function. This can be easily verified by calculating the first derivative

$$\partial_x f(x) = 1 - \frac{1}{x} \quad (\text{A3})$$

and the second one

$$\partial_x^2 f(x) = \frac{1}{x^2} > 0 \quad (\text{A4})$$

We observe that  $2\Delta I(\hat{w}_i) = f(\sum_j \omega_{ij}^2 \delta_j)$ . By Jensen's inequality, we get

$$\begin{aligned} 2f\left(\sum_j \omega_{ij} \delta_j\right) &\leq \sum_j \omega_{ij}^2 2f(\delta_j) \\ &\leq \sum_j \omega_{ij}^2 2f(\min_k(\delta_k)) \\ &= 2f(\min_k(\delta_k)) \underbrace{\sum_j \omega_{ij}^2}_{=1} \end{aligned} \quad (\text{A5})$$

Note, we used here that  $f(x)$  has its minimum at  $x = 1$  and that the eigenvalues of  $D_o$  are between 0 and 1, that is, the smaller eigenvalues maximize  $f(\delta_i)$ . This way we got an upper bound reached for  $\omega_i = \delta_{0i}$ , where  $v_0$  is the eigenvector corresponding to the smallest eigenvalue  $\delta_0 := \min_k(\delta_k)$ .  $\square$

Doing data compression by just considering the smallest eigenvalues of  $D_o$  will be found to be the right choice when considering  $\langle \Delta \text{KL}_{o,c}(\hat{r}_i) \rangle_{P(m_o)}$  as a loss function in the section A. This gives the expected loss for the expected mean of  $m_o = 0$  under  $\mathcal{P}(m_o) = \mathcal{P}(m_o | S, R_o, N_o)$ .

## Appendix B: 1D Wiener Filter Data Compression

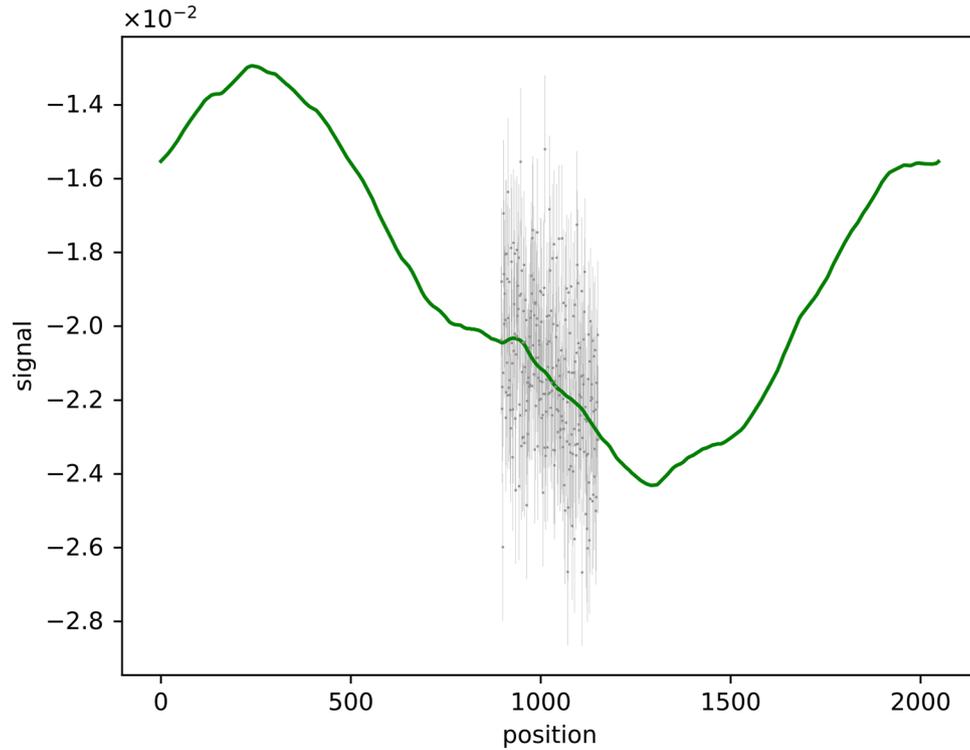
In Section 4.1, we applied our data compression method to synthetic data in the context of the generalized Wiener filter with a linear measurement equation. In this section, we are going to investigate the shape of the eigenfunctions corresponding to the eigenproblem of Equation (44) in this setting.

Therefore consider an easy set up without varying noise nor a complex mask. To ensure a certain definition, we choose the signal space to be a 1D regular grid with 2048 lattice points in one dimension. The synthetic signal and corresponding synthetic data are drawn from the prior specified in Section 4.1. The data is masked, such that only the central 256 pixels are measured. Those data are then compressed to four data points, from which the signal is inferred in a last step.

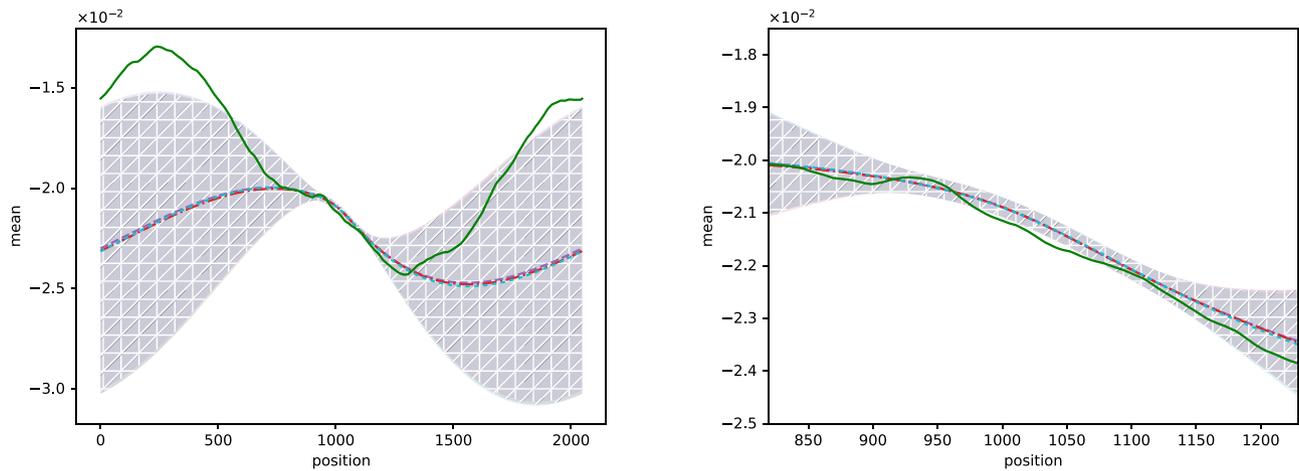
The synthetic signal and data are computed as before. However, the noise standard deviation now is constantly  $0.2 \times 10^{-2}$  and the response is set to be a mask measuring pixels 896–1152 leading to a transparent window of 256 pixels in the center of the grid. The measurement setup with signal mean, synthetic signal and data can be seen in **Figure B1**.

The consequential mean and uncertainty for the inference with the original data and the ones with the compressed data are plotted together with the ground truth in **Figure B2**. The original data has been compressed from 256 to 4 data points.

Now let us have a closer look at the eigenvectors plotted in **Figure B3**, which correspond to a back projection with  $R_c$  of the corresponding single data point being one and all the others being zero. These functions remind of Chebyshev polynomials of the first kind. The Chebyshev polynomials were fitted to the eigenvectors minimizing the mean squared error and are plotted in the same figure as the eigenvectors. One can clearly see their similarity. The lower order polynomials fit the best, while higher order polynomials deviate especially at the edges. This hints at BDC transferring the compression problem to a polynomial fit. Then the compressed data points are the amplitudes of



**Figure B1.** 1D synthetic data setup to investigate BDC. The synthetic signal is marked in green and the measured data in gray.



**Figure B2.** The same as Figure 2, but for the setup shown in Figure B1. In this simple set up original and compressed reconstruction almost perfectly coincide.

the polynomials while the compressed response stores their individual shapes.

An analytical analysis is done in the next section.

### Appendix C: Analytical Solution of the 1D Wiener Filter Data Compression

In this section, we will derive the eigenfunctions of (44) analytically for some signal on a 1D line covered by a mask of length  $L$  starting at  $x = 0$ . The response of the linear measurement equa-

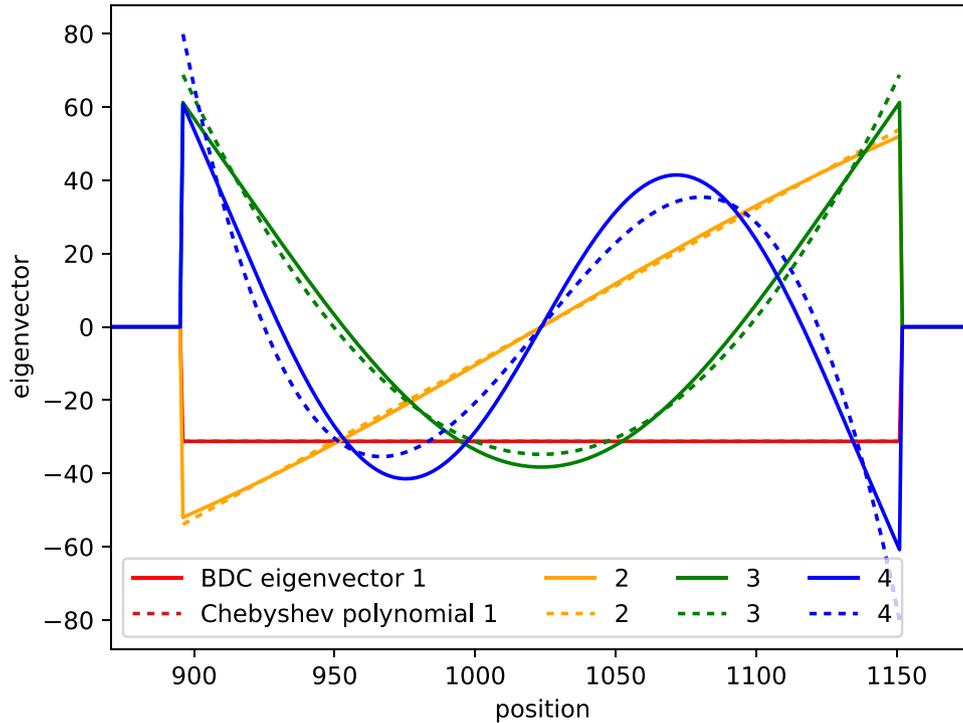
tion is

$$R_{o_{xx'}} = \delta(x - x')\chi_{[0,L]}(x'),$$

$$\text{with } \chi_{[0,L]}(x) := \begin{cases} 1 & \text{for } x \in [0, L] \\ 0 & \text{else} \end{cases} \quad (\text{C1})$$

The Gaussian noise  $n_o$  in the measured area has the covariance

$$N_{o_{xx'}} = n_{\text{const}}\delta(x - x') \quad (\text{C2})$$



**Figure B3.** The first four eigenvectors (solid lines) as in Figure 3, but for the setup shown in Figure B1 together with the first four Chebyshev polynomials (dashed lines). Their coincidence reveals the operating mode of BDC using polynomials fits to compress the information about the signal. Then the compressed data represent the amplitude of those polynomials and the compressed response stores their shape.

Having a look on the eigenvalue problem (44)

$$M_o S r_i = \mu_i^2 r_i \quad (C3)$$

with

$$M_o = R_o^\dagger N_o R_o \quad (C4)$$

we see, that  $r_i(x) = 0$  for  $x \notin [0, L]$  and  $\mu_i \neq 0$ . For  $x \in [0, L]$

$$\begin{aligned} S r_i &= \frac{\mu_i^2}{n_{\text{const}}} r_i \\ &=: \lambda_i r_i \end{aligned} \quad (C5)$$

We specified  $S$  by a falling power spectrum following a power law with spectral index of  $-2\alpha$  in Hartley space.

$$\begin{aligned} S &= \mathbb{H}P(|k|)\mathbb{H}^\dagger \\ &= \left(\mathbb{H}\frac{1}{k^\alpha}\right)^{1+\dagger} \\ &= \left(\Delta^{-\frac{\alpha}{2}}\right)^\dagger \Delta^{-\frac{\alpha}{2}} \end{aligned} \quad (C6)$$

with Hartley transform

$$\mathbb{H}_{,xk} = \frac{1}{\sqrt{2\pi}} \int dk [\cos(kx) + \sin(kx)] \quad (C7)$$

and Laplace operator

$$\Delta := \nabla^2 = \sum_i \partial_{x_i}^2 \quad (C8)$$

Then

$$\begin{aligned} \lambda_i r_i &= S r_i \\ &= \Delta^{-\alpha} r_i \\ &= \lambda_{\Delta_i}^{-\alpha} r_i \end{aligned} \quad (C9)$$

with eigenfunction  $r_i$  and eigenvalue  $\lambda_{\Delta_i}$  of the Laplace operator. This is equivalent to the Helmholtz equation with opposite sign. Its eigenfunctions in 2D are the Bessel functions. In one dimension with  $\alpha = 2$  as in Section 4.1, the covariance operator becomes  $S = \partial_x^{-4}$ , thus

$$\partial_x^{-4} r_i = \lambda_{\Delta_i}^{-2} r_i \quad (C10)$$

The square of the eigenvalue ensures the eigenvalues of the prior covariance to be positive. Since  $\lambda_i = \lambda_{\Delta_i}^{-\alpha}$ ,  $\lambda_i$  does not become zero.

The solution to this problem are super positions of exponential functions of the same eigenvalue

$$r_i(x) = a e^{+\sqrt{\lambda_i} x} + b e^{-\sqrt{\lambda_i} x} \quad (C11)$$

such as  $\cosh(\sqrt{\lambda_i}x)$  and  $\sinh(\sqrt{\lambda_i}x)$  for positive  $\lambda_i$ , as well as  $\cos(\sqrt{-\lambda_i}x)$  and  $\sin(\sqrt{-\lambda_i}x)$  for negative  $\lambda_i$ . For eigenvalue  $\lambda_{\Delta_i} = 0$ , also polynomials up to third order are eigenfunctions to the Laplace operator. Those belong to the largest eigenvalues of  $S$ , which are also the most informative ones according to Equation (40). This explains the proximity of the eigenvectors to Chebyshev polynomials as observed in Figure B3 for a signal power spectrum that asymptotically follows  $k^{-4}$ .

## Acknowledgements

The authors thank their reviewers for extensive and constructive feedback that helped to improve the paper a lot. T.E. and J.H.-K. acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311. P.A. acknowledges financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT). The authors thank Philipp Frank for help with the coding.

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

Bayesian statistics, data compression, Gaussian likelihood, information theory, lossy compression, signal reconstruction

Received: October 17, 2020

Revised: December 25, 2020

Published online: February 8, 2021

- 
- [1] L. Kong, Z. Liu, J. Wu, *J. Cleaner Prod.* **2020**, 273, 123142.  
 [2] How search organizes information, <https://www.google.com/search/howsearchworks/crawling-indexing/> (accessed: November 2020).  
 [3] V. Marx, *Nature* **2013**, 498, 255.

- [4] K. Grainge, B. Alachkar, S. Amy, D. Barbosa, M. Bommineni, P. Boven, R. Braddock, J. Davis, P. Diwakar, V. Francis, R. Gabrielczyk, R. Gamatham, S. Garrington, T. Gibbon, D. Gozzard, S. Gregory, Y. Guo, Y. Gupta, J. Hammond, D. Hindley, U. Horn, R. Hughes-Jones, M. Hussey, S. Lloyd, S. Mammen, S. Miteff, V. Mohile, J. Muller, S. Natarajan, J. Nicholls, et al., *Astron. Rep.* **2017**, 61, 288.  
 [5] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience Publication, Hoboken, NJ **2006**.  
 [6] K. Karhunen, *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*, Annales Academiae Scientiarum Fennicae: Ser. A. 1. Sana **1947**.  
 [7] M. Loève, *Processus stochastique et mouvement Brownien* **1948**, 366.  
 [8] D. D. Kosambi, *J. Indian Math. Soc.* **1943**, 7, 76.  
 [9] K. Pearson, *London Edinburgh Dublin Philos. Mag. J. Sci.* **1901**, 2, 559.  
 [10] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer, New York **2010**.  
 [11] B. C. Geiger, G. Kubin, arXiv preprint, arXiv:1205.6935, **2013**.  
 [12] R. Leike, T. Enßlin, *Entropy* **2017**, 19, 402.  
 [13] L. Giraldi, O. P. Le Maître, I. Hoteit, O. M. Knio, *Comput. Stat. Data Anal.* **2018**, 124, 252.  
 [14] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, Technology Press of the Massachusetts Institute of Technology, Cambridge, MA **1949**.  
 [15] R. Lehoucq, D. Sorensen, C. Yang, *ARPACK Users' Guide (Software, Environments, Tools)*, Society for Industrial and Applied Mathematics, Philadelphia, PA **1998**.  
 [16] J. Knollmüller, T. A. Enßlin, (Preprint) arXiv:1812.04403, v1, submitted: Dec **2013**.  
 [17] J. Knollmüller, T. A. Enßlin, arXiv preprint, arXiv:1901.11033, **2020**.  
 [18] M. Vergassola, E. Villermaux, B. I. Shraiman, *Nature* **2007**, 445, 406.  
 [19] X. Cai, L. Pratley, J. McEwen, *Mon. Not. R. Astron. Soc.* **2017**, 485, 4559.  
 [20] E. Jones, T. Oliphant, P. Peterson, SciPy: Open source scientific tools for Python, <http://www.scipy.org/> (accessed: October 2020).  
 [21] P. Arras, P. Frank, R. Leike, R. Westermann, T. A. Enßlin, *A&A* **2019**, 627, A134.  
 [22] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, *J. Am. Stat. Assoc.* **2017**, 112, 859.  
 [23] S. Ananthkrishnan, *J. Astrophys. Astronomy Suppl.* **1995**, 16, 427.  
 [24] W. Raja, *Ph.D. Thesis*, Jawaharlal Nehru University, New Delhi, India **2014**.  
 [25] P. Arras, R. A. Perley, H. L. Bester, R. Leike, O. Smirnov, R. Westermann, T. A. Enßlin, arXiv preprint, arXiv:2008.11435, **2020**.  
 [26] H. Junklewitz, M. R. Bell, M. Selig, T. A. Enßlin, *A&A* **2016**, 586, A76.