Bayesian Approximate Message Passing for Distributed Compressed Sensing

Gábor Hannák

Institute of Telecommunications
Technische Universität Wien

This dissertation is submitted for the degree of
Doktor der technischen Wissenschaften

August 2017

Advisor

Univ.-Prof. Dipl.-Ing. Dr.-Ing. Norbert Görtz
Institute of Telecommunications
Technische Universität Wien
Austria

Examiners

Prof. Ph.D. Michael E. Davies
Institute of Digital Communications
The University of Edinburgh
Scotland, United Kingdom

Ao. Univ.-Prof. Dipl.-Ing. Dr. techn. Gerald Matz
Institute of Telecommunications
Technische Universität Wien
Austria

Abstract

In recent years, the low-dimensional representation of high-dimensional signals has been recognized as an essential concept in modern signal processing. An important family of problems is subsumed under the term *compressed sensing* (CS). CS copes with the reconstruction or estimation of a high-dimensional vector from a (noisy) underdetermined system of linear equations, assuming that the measured vector has only a relatively low number of nonzero components. Under mild conditions on the dimensions and the structure of the system matrix (measurement matrix), reconstruction or robust estimation is feasible. *Approximate message passing* (AMP), an approximate and highly simplified version of loopy belief propagation, has proven to cope efficiently with high-dimensional sparse problems. Its Bayesian version, *Bayesian approximate message passing* (BAMP), which is an approximate *minimum mean squared error* (MMSE) estimator, is a versatile algorithm that can incorporate prior knowledge about the measured vector in the form of a prior *probability density function* (pdf) of its components. When there is a set of measured vectors which are somehow dependent, e.g., jointly sparse (i.e., their sets of nonzero components are identical), joint recovery proves advantageous. More specifically, when a multivariate prior pdf for the vector components of the jointly measured vectors is available, BAMP can be extended to its vector version, the *vector Bayesian approximate message passing* (V-BAMP). V-BAMP is an approximate MMSE estimator for the whole set of jointly measured vectors, and its analysis can be derived from the scalar BAMP. Specifically, the *state evolution* (SE) equations provide an analytical prediction for the residual *mean squared error* (MSE) of the vector estimates. Understanding the dynamics of SE in terms of fixed points as a function of the signal prior, the noise parameters, and the sampling rate is of crucial importance because it uncovers the expected behavior of V-BAMP.

In this work we investigate the V-BAMP algorithm. In particular, both the increasing number of jointly sparse measured vectors, as well as correlation between the nonzero signal components and the noise components are explored. The SE equations are extended to the multivariate case and extensive simulations show the effect of the number of measurement vectors and the effect of having correlation on the recovery.

We show that (i) arbitrary signal and noise correlations can be eliminated in the joint measurement case using a linear transform; (ii) V-BAMP is equivariant with respect to linear transformations; and (iii) for the widely employed multivariate *Bernoulli-Gauss* (BG) signal prior the uncorrelatedness of the signal and of the noise are preserved through the V-BAMP iterations. It follows that the decorrelation transform has to be done only once before starting V-BAMP, and neither the convergence nor the MSE performance are affected. Furthermore, the analysis of V-BAMP with BG signals is reduced to the case with diagonal signal and noise covariance structure. Recently, based on the analogy between the statistical physics of large disordered systems and loopy belief propagation, the replica method was used to approximate the MMSE of the Bayesian estimator of the CS measurement, for BG signal prior with standard Gaussian nonzero signal components and isotropic uncorrelated Gaussian noise. In this work, the replica analysis is extended to the case with arbitrary (anisotropic) uncorrelated Gaussian noise. Together with the joint decorrelation transform and the equivariance property of V-BAMP, the replica analysis turns out to predict the dynamics of V-BAMP for BG signals and Gaussian noise, with arbitrary signal and noise correlations. Simulations confirm the analogy between the SE analysis and the replica analysis, and demonstrate the effect of measuring multiple signals and that of signal correlation from many aspects.

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the references and acknowledgements.

Gábor Hannák
August 2017

Acknowledgements

I wish to express my sincere gratitude to Norbert Görtz, Michael (Mike) E. Davies, Gerald Matz, and Martin Mayer, without whom it would not have been possible to complete this work. I also thank Georg Pichler, Peter Berger, Alessandro Perelli, and Osman Musa for their professional and motivational support.

Table of Contents

Notation and Definitions

deterministic scalar, column vector, and matrix	$a, \mathbf{a}, \mathbf{A}$
random scalar, column vector, and matrix	$\mathsf{a}, \mathbf{a}, \mathbf{A}$
nth component of vector	a_n
columns of matrix	$\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_N)$
(m, n)th entry of a matrix	$(\mathbf{A})_{m,n} = A_{m,n}$
vector and matrix transpose	$\mathbf{a}^T, \mathbf{A}^T$
matrix inverse	$\mathbf{A}^{-1}$
matrix determinant	$\lvert\mathbf{A}\rvert$
$N \times N$ identity matrix	$\mathbf{I}_N$
$M \times N$ all zeros matrix	$\mathbf{0}_{M\times N}$
diagonal matrix with entries on the diagonal given by vector or ordered list	$\mathrm{diag}(\mathbf{a})$
vector of diagonal entries of a matrix	$\mathrm{diag}(\mathbf{A})$
outer product of a (column) vector with itself	$\langle\mathbf{a}\rangle = \mathbf{a}\mathbf{a}^T$
vector composed of the identically indexed components from an ordered set of B vectors	$\vec{\mathbf{a}}_n = (a_n(1), \ldots, a_n(B))^T$
vector p-norm ($p \geq 1$)	$\lVert\mathbf{a}\rVert_p$
matrix Frobenius norm	$\lVert\mathbf{A}\rVert_F$
(ordered) set	$\mathcal{S}$
cardinality of a set	$\lvert\mathcal{S}\rvert$
set of positive integers up to N	$[N] = \{1, \ldots, N\}$
vector with components indexed by set	$\mathbf{a}_\mathcal{S}$
matrix with columns indexed by set	$\mathbf{A}_\mathcal{S}$
probability of an event	$\mathrm{P}\{\cdot\}$
expectation of a random quantity	$\mathrm{E}\{\cdot\}$

variance of a random quantity	$\mathrm{Var}\{\cdot\}$
sample variance of a quantity	$\mathrm{Var}(\cdot)$
covariance of a random vector	$\mathrm{Cov}\{\cdot\}$
sample covariance of a vector	$\mathrm{Cov}(\cdot)$
discrete uniform distribution	$\mathcal{U}[\ ]$
(multivariate) normal distribution with mean $\boldsymbol{\mu}$ and (co-)variance $\boldsymbol{\Sigma}$	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
(multivariate) normal distribution function with mean $\boldsymbol{\mu}$ and (co-)variance $\boldsymbol{\Sigma}$ (evaluated at $\mathbf{x}$)	$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$
iteration index (e.g., in an algorithm)	$(\cdot)^{(t)}$
convolution of two functions	$*$

Sets

- Support (set): for a vector $\mathbf{a}$,

$$\mathrm{supp}(\mathbf{a}) = \{n \,:\, a_n \neq 0\}\,.$$

- For a column vector $\mathbf{a}$ of dimension N and set $\mathcal{S} = \{n_1, \ldots, n_{|\mathcal{S}|}\} \subseteq [N]$,

$$\mathbf{a}_{\mathcal{S}} = \mathbf{F}\mathbf{a} \;\text{ with } \mathbf{F} \in \{0,1\}^{|\mathcal{S}| \times N} \text{ and } F_{i,n_i} = 1 \;\Leftrightarrow n_i \in \mathcal{S}\,,$$

  i.e., $\mathbf{a}_{\mathcal{S}}$ is the vector composed of the components of $\mathbf{a}$ whose indices are in $\mathcal{S}$ (preserving the order).

- For a matrix $\mathbf{A}$ of dimension $M \times N$ and set $\mathcal{S} = \{n_1, \ldots, n_{|\mathcal{S}|}\} \subseteq [N]$,

$$\mathbf{A}_{\mathcal{S}} = \mathbf{A}\mathbf{F} \text{ with } \mathbf{F} \in \{0,1\}^{|\mathcal{S}| \times N} \text{ and } F_{i,n_i} = 1 \;\Leftrightarrow n_i \in \mathcal{S}\,,$$

  i.e., $\mathbf{A}_{\mathcal{S}}$ is the matrix composed of the columns of $\mathbf{A}$ whose indices are in $\mathcal{S}$ (preserving the order).

Probabilities

- Discrete uniform distribution: for a finite set $\mathcal{S} = \{s_1, \ldots, s_N\}$, the discrete uniform distribution $\mathcal{U}[s_1, \ldots, s_N]$ is defined by the (generalized) probability

density function

$$f_{\mathsf{s}}(s) = \sum_{n=1}^{N} \frac{1}{N}\delta(s - s_n) \quad \text{or} \quad P\{\mathsf{s} = s_n\} = \frac{1}{N} \quad \forall n\,.$$

Functionals

- Dirac delta (generalized) function: in the strict sense the Dirac delta is not a function, but defined by the integral

$$f(a) = \int_{\mathcal{R}(f)} f(x)\delta(x - a)dx$$

over any function $f : \mathcal{R}(f) \to \mathcal{I}(f)$ with $\mathcal{R}(f), \mathcal{I}(f) \subseteq \mathbb{R}$.

- Multivariate Dirac delta (generalized) function: for all functions $f : \mathcal{R}(f) \to \mathcal{I}(f)$ with $\mathcal{R}(f), \mathcal{I}(f) \subseteq \mathbb{R}^N$ the N-D Dirac delta satisfies

$$f(\mathbf{a}) = \int_{\mathcal{R}(f)} f(\mathbf{x})\delta(\mathbf{x} - \mathbf{a})d\mathbf{x}\,.$$

Miscellaneous

- *Mean squared error* (MSE): the MSE between two vectors of dimension N is defined as

$$\mathrm{MSE}(\mathbf{a}, \mathbf{b}) = \frac{1}{N}\|\mathbf{a} - \mathbf{b}\|_2^2\,.$$

- Decibel notation: quantities $x \in \mathbb{R}$ in dB units are defined as

$$x\,\mathrm{dB} = 10^{\frac{x}{10}}\,.$$

Chapter 1

Introduction

As humanity is ever faster developing and forming its environment to its own advantage, it seems unavoidable that information technology and digitalization will dominate the upcoming era [1, 2]. The progress of hardware development [3] resulted in both the explosion of available processing power and massively available (affordable) hardware devices. These devices not only surround and aid our everyday lives, but by assigning to them ever more serious and complicated tasks, ranging from a digital calendar to controlling a city's transportation and energy supply system, humans are bound to them stronger than ever before. It is the task of signal processing scientists to exploit the potential that arises under these circumstances: the responsible design of efficient acquisition, storage, and processing methods for the next generation apparatus.

An important paradigm in signal processing is the fact that the captured data carries redundancy. In particular, it is possible to compress a bundle of data such that it can be restored in its whole, or, in some cases, the valuable pieces of contained information can be retrieved from the compressed data, when necessary. Prime examples of this phenomenon are image [4] and audio compression [5, 6]. *Compressed sensing* (CS) [7–10] relies on the observation that the two-step process of *data acquisition* or *measurement* followed by *data compression* is suboptimal. More specifically, the data compression step corrects for the suboptimality of the data acquisition phase which preserved the unnecessary redundancy.
Based on this consideration, CS aims at replacing this two-step procedure by a single step, which naturally captures the data in a compressed form. Take the following historical example: in a large population, say of more than thousand individuals, only a small fraction possesses a medical condition that renders the individual *defective* for some purpose, e.g., military service. By blood test, it is possible to test a person for

Fig. 1.1 Original highly redundant image and its reconstruction with 2% respectively 5% measurements.

its condition. Testing all individuals one by one is a tedious and expensive process. However, forming groups of individuals and performing single tests first on the groups (by mixing fractions of the individuals' blood samples), one can, with high probability, discard large healthy fractions of the population before continuing to perform individual tests within the groups that turn out to contain at least one defective member. When the portion of defective members is relatively low, e.g., $< 30\%$, this procedure makes the detection much more efficient than individual testing. This observation by Robert Dorfman in 1943 [11] laid the foundation of the branch of mathematics called *group testing*.

In 2006, Duarte et al. created a proof of concept setup for a real-world CS application: the *single-pixel camera* [12] is capable of acquiring an optical image loaded with redundancy with far less measurements than the number of pixels would suggest. Typically, in order to obtain the raw data, the number of measurements is the number of pixels on the camera sensor. After acquiring every pixel value, the raw image data is compressed in order to obtain a version optimized for storage and human view experience. With the single pixel camera, the researchers could control how much information is sequentially acquired through the optics. Figure 1.1 shows the ideal 256×256 image (acquired by performing $65536 = 100\%$ measurements), and the images obtained after sensing with the single pixel camera with only 2% respectively 5% measurements (taken from [12]). The original image is very redundant and, clearly, the carried information is contained after strong compressive measurement.

In this chapter we outline the mathematical foundations of CS, reconstruction methods, and some generalizations. First, the concept of sparsity is introduced and we discuss how it arises in a class of redundant signals. Then, conditions for the unique mapping and robust sampling of sparse signals are discussed. A short overview of existing recovery methods follows, and the chapter is closed with a short list of possible generalizations of the classical CS problem.

1.1 Sparsity and the Linear Inverse Problem

1.1.1 The Linear Inverse Problem

At the very core of applied linear algebra lies the problem of solving the system of linear equations [13]

$$\mathbf{y} = \mathbf{A}\mathbf{x}\,, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^M$ and $\mathbf{A} \in \mathbb{R}^{M\times N}$ are known and $\mathbf{A}$ has full column rank. The (cardinality of the) solution set $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}\,|\mathbf{A}\hat{\mathbf{x}} = \mathbf{y}\}$ depends on the dimensions M, N:

1. If $M > N$ the system is overdetermined and has no solution: $\hat{\mathcal{X}} = \emptyset$.
2. If $M = N$ the system has one unique solution: $\hat{\mathcal{X}} = \{\mathbf{A}^{-1}\mathbf{y}\}$.
3. If $M < N$ the system is underdetermined and has infinitely many solutions: $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}\,|\hat{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} + \mathbf{v}\,, \mathbf{A}\mathbf{v} = \mathbf{0}\}$.

If the system is underdetermined, the Moore-Penrose pseudoinverse $\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ and the associated solution $\hat{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y}$ has a special role: it is the solution with minimum ℓ_2-norm. Note that the mapping between $\mathbf{x}$ and $\mathbf{y}$ is not unique, i.e., infinitely many vectors of dimension N result in $\mathbf{y}$. In general, the linear inverse problem consists of finding the solution of (1.1) when the unknown is subject to some nonlinear constraint(s).

1.1.2 Sparsity

Consider a signal vector $\mathbf{s} \in \mathbb{R}^N$, which carries redundancy in some sense. To formulate this in a mathematical way, we assume that the vector has only a few nonzero coefficients in some orthonormal basis $\mathbf{B} \in \mathbb{R}^{N\times N}$, i.e., its representation coefficient vector

$$\mathbf{x} = \mathbf{B}\mathbf{s}$$

has many zeros.

Definition 1 *A vector* $\mathbf{x}$ *is* K*-sparse if it has at most* K *nonzero components.*

Example: In Figure 1.2 the original *cameraman* image (of size $N = 256\times 256$ pixels) was transformed using the discrete cosine basis [14]. The smallest 90% of the coefficients (in magnitude) were discarded (the threshold is represented by the red dashed line), which resulted in a coefficient vector with sparsity ratio $1 - K/N = 0.9$. The compressed

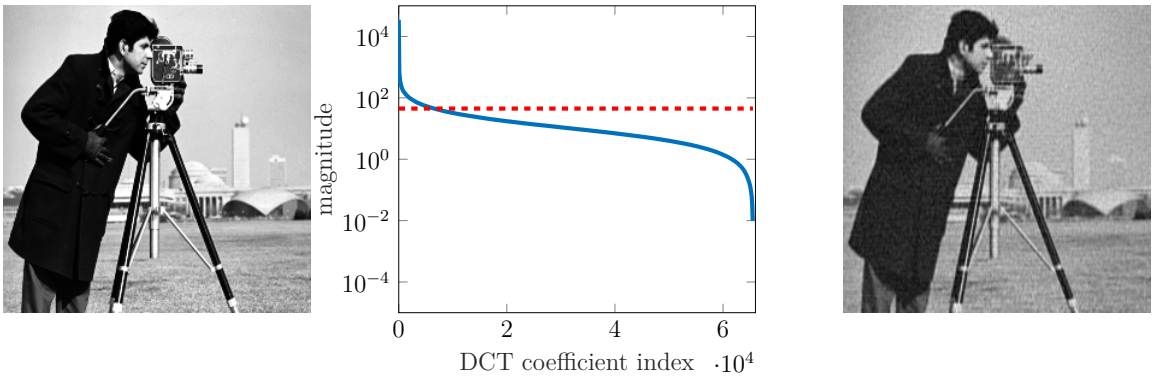


Fig. 1.2 Original 256×256 image, its transform coefficients sorted by magnitude, and a compressed image restored from 10% of the coefficients.

image after inverse transformation can be seen on the right. Clearly, there is visible quality loss, but the main content and most details of the image are preserved. Thus, the image is compressible and can be represented by a sparse vector.
Another useful way to capture sparsity is the ℓ_0-seminorm

$$\|\mathbf{x}\|_0 = \lim_{p \to 0} \left(\sum_{n=1}^{N} |x_n|^p \right)^{1/p} = |\mathrm{supp}(\mathbf{x})| \,,$$

which counts the number of components in $\mathbf{x}$ that are nonzero. Then, the fact that a vector $\mathbf{x}$ is K-sparse can be compactly written as $\|\mathbf{x}\|_0 \leq K$.

1.1.3 Unique Mapping of Sparse Vectors

If $\mathbf{x} \in \mathbb{R}^N$ is sparse, then under certain constraints the *measurement* (mapping)

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

with $\mathbf{A} \in \mathbb{R}^{M \times N}$ is unique even if the system is underdetermined, i.e., if $M < N$. A simple condition on the measurement matrix can be easily shown: in order to map each pair of distinct K-sparse vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ to different measurements $\mathbf{y}_1 \neq \mathbf{y}_2$, one needs $\mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) \neq \mathbf{0}$. This is guaranteed if every set of $2K$ columns in $\mathbf{A}$ is linearly independent.

Definition 2 *The spark of a matrix* $\mathbf{A}$ *is the smallest number* n *such that there exist* n *columns in* $\mathbf{A}$ *that are linearly dependent.*

Corollary 1 *If* $\mathrm{spark}(\mathbf{A}) > 2K$*, the mapping* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *is unique for all vectors* $\mathbf{x}$ *with* $\|\mathbf{x}\|_0 \leq K$ *[15].*

Definition 3 *The original CS problem attempts to find the sparsest vector* $\hat{\mathbf{x}}$ *that is consistent with the noiseless measurement* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *(or the noisy measurement* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$*.)*

The straightforward approach to solve the original CS problem is ℓ_0-minimization

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_0 \text{ s.t. } \mathbf{y} = \mathbf{A}\tilde{\mathbf{x}}\,, \tag{1.2}$$

where, if $\|\mathbf{x}\|_0 \leq K$ and $\operatorname{spark}(\mathbf{A}) > 2K$, $\hat{\mathbf{x}} = \mathbf{x}$. Unfortunately, to compute the spark of a matrix is NP-hard [16]. Moreover, solving the optimization problem (1.2) is combinatorially hard because it involves an exhaustive search through all column combinations of $\mathbf{A}$ of size K.

1.1.4 Robust Sampling

In practice, oftentimes the measurement does not perfectly match the mathematical model. Thus, the noisy measurement model is introduced in the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}\,, \tag{1.3}$$

where $\mathbf{w}$ captures measurement noise and modelling errors. Since typically the additive noise is unknown and only described statistically, one becomes interested in *robust sampling*. This requires that the distance between a pair of K-sparse vectors is approximately preserved by the mapping through $\mathbf{A}$. The rationale behind this is that if the minimum distance between the images of sparse vectors exceeds the noise level *with high probability* (whp), the chance of recovering the wrong sparse vector is minimal.

Definition 4 [17] *The matrix* $\mathbf{A} \in \mathbb{R}^{M \times N}$ *fulfills the restricted isometry property (RIP) of order* s *with RIP constant* δ_s *if for all* s*-sparse vectors* $\mathbf{v} \in \mathbb{R}^N$

$$(1 - \delta_s)\|\mathbf{v}\|_2^2 \leq \|\mathbf{A}\mathbf{v}\|_2^2 \leq (1 + \delta_s)\|\mathbf{v}\|_2^2\,.$$

In order to achieve robust sampling, one is interested in a low RIP constant $\delta_s = \delta_{2K}$, so that the measurements of pairs of $2K$-sparse vectors are well separated. The RIP constant is bounded from below by the scaling of the dimensions $N, M,$ and K. For example, it can be shown that for a measurement matrix $\mathbf{A}$ that fulfills the RIP with δ_{2K} such that $\delta_{2K} \leq \frac{1}{2}$, necessarily $M \geq 0.3K \log N/K$ [10, Theorem. 1.4]. Unfortunately,

for a given matrix $\mathbf{A}$ and a RIP constant δ_s, it is NP-hard to decide whether $\mathbf{A}$ fulfills the RIP with δ_s, as well as to determine the minimum RIP constant. Furthermore, it is nearly impossible to construct deterministic measurement matrices with desired dimensions and RIP constant [18].

In contrast, probability theory provides CS with very promising results. In particular, random constructions deliver measurement matrices that possess the desired properties with overwhelming probability. The interested reader is referred to [10, 16, 18, 19] for comprehensive material on random matrix constructions and their properties. The most relevant cases are the *Bernoulli* or *Rademacher measurement matrix* and the *Gaussian measurement matrix*, whose entries are *independent and identically distributed* (i.i.d.) zero-mean discrete uniform respectively Gaussian, i.e.,

$$A_{m,n} \sim \mathcal{U}\left[-\frac{1}{\sqrt{M}}, \frac{1}{\sqrt{M}}\right] \text{ respectively } A_{m,n} \sim \mathcal{N}\left(0, \frac{1}{M}\right),$$

and they obey normalized columns. These two matrices belong to the class of *sub-Gaussian matrices* [20], for which the following theorem holds.

Theorem 1 [19, Theroem 9.2] *Let* $\mathbf{A}$ *be an* $M \times N$ *sub-Gaussian random matrix with normalized columns. Then there exists a constant* $C > 0$ *(independent of* M, N, δ_s*) such that the RIP constant of* $\mathbf{A}$ *satisfies* $\delta_s \leq \delta$ *with probability at least* $1 - \epsilon$ *provided* $M \geq 2C\delta^{-2}(s \ln(eN/s) - \ln(2\epsilon))$.

Setting $\epsilon = 2\exp(-\delta^2 M/(2C))$ yields the condition $M \geq 2C\delta^{-2} s \ln(eN/2)$ with probability $1 - 2\exp(-\delta^2 M/(2C))$.

Literature supports the wide range of applicability of both the Bernoulli and the Gaussian matrix: In *radio-frequency identification* (RFID) [21] and multiuser communication systems [22] nodes can be identified by a unique binary signature sequence. When chosen in a random fashion, the sequences form columns of a Bernoulli matrix, which in the communication system model forms a valid measurement matrix. When the node activity is sparse, i.e., only a small fraction of the nodes is transmitting at the same time, activity detection corresponds to a valid CS measurement. The Gaussian measurement matrix has applications, e.g., in radar imaging [7] and compressive analog sampling [23]. For a comprehensive survey on CS applications the interested reader is referred to [24] and Chapter 4.

1.2 Reconstruction Methods

In theory, $M \geq 2K$ measurements are sufficient to reconstruct any K-sparse vector from noiseless measurements [19, Theorem 2.14]: for every $N \geq 2K$ there exists a measurement matrix $\mathbf{A} \in \mathbb{R}^{2K \times N}$ such that every K-sparse vector $\mathbf{x}$ can be reconstructed from the measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$ via

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_0 \text{ s.t. } \mathbf{y} = \mathbf{A}\tilde{\mathbf{x}}\,. \tag{1.4}$$

(1.4) is referred to as ℓ_0-*minimization* problem, and it can be proven to be NP-hard since solving it involves an exhaustive search through all column K-combinations of $\mathbf{A}$. Moreover, the *noisy* ℓ_0-*minimization*

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_0 \text{ s.t. } \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 \leq \eta\,, \tag{1.5}$$

which aims at estimating the solution of (1.3) with any nonnegative η, is NP-hard as well [19, Theorem 2.17].

1.2.1 Convex Relaxation

The ℓ_0-minimization (1.4) is NP-hard. Observing that $\|\mathbf{z}\|_p^p$ tends to $\|\mathbf{z}\|_0$ as $p \to 0$, a sequence of approximations of the solution can be obtained by

$$\hat{\mathbf{x}}_{(p)} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_p \text{ s.t. } \mathbf{y} = \mathbf{A}\tilde{\mathbf{x}}$$

for $p > 0$. The smallest p for which this problem becomes convex is $p = 1$ and the corresponding problem is referred to as ℓ_1-*minimization* or *basis pursuit (BP)*:

$$\hat{\mathbf{x}}_{(1)} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_1 \text{ s.t. } \mathbf{y} = \mathbf{A}\tilde{\mathbf{x}}\,. \tag{1.6}$$

Under mild conditions, the solution of BP and that of ℓ_0-minimization are identical and thus the CS reconstruction problem can be replaced by a convex optimization problem, for which there is a variety of methods. Analogously to (1.5), it is possible to incorporate a quadratic constraint in order to account for measurement and model inaccuracies, and state the *quadratically constrained BP*:

$$\hat{\mathbf{x}}_{\mathrm{BP}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_1 \text{ s.t. } \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 \leq \eta\,. \tag{1.7}$$

The solution of (1.7) is strongly related to the dual problem termed *BP denoising*:

$$\hat{\mathbf{x}}_{\text{BPDN}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 \text{ s.t. } \|\tilde{\mathbf{x}}\|_1 \leq \lambda \,,$$

with a parameter $\lambda > 0$ that trades off the sparsity and the empirical ℓ_2 error. Both the quadratically constrained BP and BP denoising are strongly related to the solution of the *least absolute shrinkage and selection operator* (LASSO) [25]

$$\hat{\mathbf{x}}_{\text{lasso}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_1 \,, \tag{1.8}$$

where the parameter λ, again, trades off the sparsity and the empirical ℓ_2 error. For an overview of existing convex optimization methods developed for CS and in particular BP and LASSO the interested reader is referred to [26, 27]. These methods' shortcomings become apparent when the dimensions get large. In many applications, the CS equation's dimensions reach the range of hundreds of thousands or even millions, where convex optimization methods can become slow.

1.2.2 Greedy Algorithms

Greedy algorithms are iterative algorithms that are based on variations of the following procedure:

1. Start with the empty set as the support estimate of the unknown $\mathbf{x}$.
2. Search for the column in the measurement matrix $\mathbf{A}$ (that has not been selected previously) which, together with the previously selected columns, best explains the measurement $\mathbf{y}$.
3. Add the selected column's index to the support set estimate and repeat from 2.

The two most prominent greedy algorithms are the *orthogonal matching pursuit* [28] and the *compressive sampling matching pursuit* [29]. Because typically every iteration in the algorithm involves solving a least squares problem with the matrix $\mathbf{A}$, the applicability of greedy methods is limited by the problem dimensionality.

1.2.3 Iterative Thresholding Algorithms

By multiplying the measurement equation (1.3) with the adjoint measurement matrix from the left one can approximately invert the measurement:

$$\begin{aligned}\mathbf{A}^T\mathbf{y} &= \mathbf{A}^T\mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{w} \\ &= \mathbf{x} + \underbrace{(\mathbf{A}^T\mathbf{A} - \mathbf{I})\mathbf{x} + \mathbf{A}^T\mathbf{w}}_{\text{additive noise}}\end{aligned}$$

If the measurement matrix design is suitable, $\mathbf{A}^T\mathbf{A}$ is close to being an identity matrix. It follows that the norm of $(\mathbf{A}^T\mathbf{A} - \mathbf{I})\mathbf{x}$ is small compared to that of $\mathbf{x}$ and can be interpreted as noise (as the two are also uncorrelated). Thresholding reduces the additive noise part in an iterative manner, where the threshold is based on either prior knowledge of the sparsity $N - K$, or the noise statistics. It has been shown that *iterative hard thresholding* [30] solves the ℓ_0-regularized minimization

$$\operatorname*{argmin}_{\tilde{\mathbf{x}} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_0 \,,$$

while *iterative soft thresholding* [31] solves the ℓ_1-regularized minimization

$$\operatorname*{argmin}_{\tilde{\mathbf{x}} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_1 \,.$$

A comprehensive comparison and extensions can be found, e.g., in [32–34]. Thresholding algorithms are simple to implement, have low computational complexity, and yield reasonable recovery performance even for huge problem dimensions.

1.2.4 The Probabilistic Approach and Message Passing Algorithms

As an alternative to the deterministic sparsity concept, where K of N entries of the unknown are nonzero, the probabilistic approach assumes a prior *probability density function* (pdf) valid independently and identically on the N components:

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{n=1}^{N} f_{\mathsf{x}_n}(x_n) = \prod_{n=1}^{N} f_{\mathsf{x}}(x_n) \,. \tag{1.9}$$

For instance, a probabilistic model for the sparse vector reads

$$f_{\mathsf{x}}(x_n) = (1 - \epsilon)\delta(x_n) + \epsilon g_{\mathsf{x}}(x_n) \,,$$

where $1 \gg \epsilon > 0$ and g_{x} is the pdf of the nonzero components. Then, it is possible to define the corresponding graphical model [35] for the measurement equation (1.3) and perform approximate loopy belief propagation [36, 37]. The estimate resulting from *approximate message passing* (AMP) [38] is closely related to the solution of the LASSO (1.8). The extended version, *Bayesian approximate message passing* (BAMP), is an approximate *minimum mean squared error* (MMSE) estimator, when the measurement instance $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$ and the prior (1.9) is given [39–41]. Variations of AMP are very simple to implement, converge fast, and have excellent recovery performance. Furthermore, they are very flexible due to the incorporation of prior knowledge. Thus, they are suitable for very high problem dimensions and a wide range of applications. This work mainly focuses on BAMP and its extensions.

1.3 Generalizations of the Linear Inverse Problem

In this section we outline some popular generalizations, special cases, and related branches of the original CS problem (cf. Definition 3).

1.3.1 Group and Block Sparsity

Suppose that the nonzero components are not distributed arbitrarily in $\mathbf{x}$ but appear in groups. Then, the index set $[N]$ can be partitioned into G groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ (possibly nonuniformly) such that $\mathbf{x}$ is sparse in a group sense, i.e., only a relatively small fraction of the groups contain nonzero components. This signal model is referred to as *group* or *block sparsity*.

Even though standard CS recovery methods can solve this problem, exploiting the prior knowledge about the group structure yields significant advantages. A wide range of standard methods have been generalized that solve this problem more efficiently (even for overlapping groups), e.g., greedy methods [42, 43], convex optimization methods [44–46], and methods based on message passing [47, 48].

1.3.2 Joint Sparsity

Say, B vectors of the same dimension are measured, whose nonzero patterns are identical, i.e.,

$$\mathbf{y}(b) = \mathbf{A}(b)\mathbf{x}(b) + \mathbf{w}(b)\,, \quad b = 1, \ldots, B\,, \tag{1.10}$$

with $\mathrm{supp}(\mathbf{x}(b')) = \mathrm{supp}(\mathbf{x}(b))$, $b, b' \in [B]$ [49, 50]. An important distinction can be made: if the B measurement matrices are identical, the setup is referred to as the *multiple measurement vectors* (MMV) problem, and otherwise as the *distributed compressed sensing* (DCS) problem. In both cases the naive approach is to perform B individual recoveries and combine the results such that the support sets match, which is clearly suboptimal. Better recovery performance is achieved when the knowledge of the joint sparsity is exploited beforehand and *joint recovery* is performed. A collection of methods [51–54], some tailored for specific applications, has been elaborated, e.g., for (medical) imaging [55–58], direction of arrival estimation [59], RFID [60], and multiuser communications [61, 62].

1.4 Contribution and Outline

In this work we describe the BAMP algorithm and its multivariate extension, the *vector Bayesian approximate message passing* (V-BAMP), for the MMV and the DCS scenarios that inherently copes with joint sparsity, and arbitrary signal and arbitrary noise correlations. We present the *state evolution* (SE) equations for the multivariate case and show through extensive numerical simulations how multiple measured vectors and signal correlation affect the recovery relative to the scalar case, and give insight into the dynamics of V-BAMP in terms of the SE equations and the *phase transition* (PT) property. We demonstrate that the PT is only present in the CS regime, i.e., when the unknown signal is sparse. We prove for the jointly sparse CS measurement that arbitrary signal and noise correlations can be eliminated. In particular, the measurement equation is transformed using an invertible linear transform, thereby obtaining an equivalent measurement model with purely diagonal signal and noise structure. Moreover, we prove that V-BAMP and its SE are equivariant with respect to such transformations, i.e., the dynamics and performance of V-BAMP are not affected by the joint decorrelation. We show that for signals with multivariate *Bernoulli-Gauss* (BG) prior, the diagonal signal and noise statistics are preserved through V-BAMP iterations. Thus, in terms of analysis, the set of all measurement instances with BG prior are reduced to the set of those with purely diagonal correlation structure. It follows that the state of the B-dimensional V-BAMP for the BG prior is not $B(B+1)/2$ dimensional, but only B dimensional. Furthermore, for the jointly diagonal CS measurement with BG signals we derive the estimation MMSE using the replica method, whose dynamics matches with that of the SE. In particular, the local maxima of the free energy function correspond to stable fixed points of the SE equation. We hypothesize that V-BAMP

is a gradient ascent algorithm on the free energy function. Using the both the SE and the replica analysis we show that when the number of jointly sparse vectors becomes large, V-BAMP is not characterized by a PT anymore, but the estimation *mean squared error* (MSE) decreases smoothly with the sampling rate. Finally, we discuss a number of modern applications of CS with emphasis on the potential of V-BAMP.

In Chapter 2, we introduce the scalar BAMP algorithm and discuss its variables and properties, and specify two signal priors of interest, the BG prior and the discrete prior. Then, we demonstrate the analytical properties of BAMP in terms of the SE equation and the PT.
In Chapter 3, we briefly discuss the differences between the DCS and the MMV scenarios, and present the V-BAMP algorithm along with an examination of its properties. Then, signal priors of interest and reestimation/exploitation of soft information are explored. Next, we discuss the multivariate SE equations and the PT for the multivariate V-BAMP with the BG prior. The joint decorrelation procedure is followed by a discussion on correlated CS, and we close the chapter with the replica analysis for CS with signals with BG prior.
In Chapter 4, three applications are briefly outlined: RFID, activity detection and channel estimation, and *quadrature amplitude modulation* (QAM) demodulation in wireless multiuser communication systems.
We conclude this work in Chapter 5.

Chapter 2

Bayesian Approximate Message Passing

The probabilistic approach to CS assumes both the unknown signal vector and the additive noise vector to be a realization of a random vector. Thus, the measurement becomes a mapping of two random vectors onto one random vector:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}\,. \tag{2.1}$$

The components of the additive noise are assumed to be i.i.d. zero-mean normal with variance σ_{w}^2, i.e.,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathsf{w}}^2 \mathbf{I}_M) \rightarrow \mathsf{w}_m \sim \mathcal{N}(0, \sigma_{\mathsf{w}}^2)\,. \tag{2.2}$$

The signal $\mathbf{x}$ is characterized by the *prior* pdf. Its components are i.i.d. with

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{n=1}^{N} f_{\mathsf{x}_n}(x_n) = \prod_{n=1}^{N} f_{\mathsf{x}}(x_n)\,. \tag{2.3}$$

The assumption of sparsity in CS can be incorporated using the multivariate Dirac delta in the prior pdf, i.e., it is assumed that

$$f_{x_n}(x_n) = f_{\mathsf{x}}(x_n) = (1-\epsilon)\delta(x_n) + \epsilon f_{nz}(x_n)$$

with $0 < \epsilon \ll 1$. Here, $f_{nz} \neq \delta$ is the distribution of x_n given that $n \in \mathrm{supp}(\mathbf{x})$. In this probabilistic setting, *exact recovery* is not possible anymore. A common approach is to search for the vector $\hat{\mathbf{x}}$ which minimizes the expected MSE:

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}} \in \mathbb{R}^N}{\operatorname{argmin}}\, \mathrm{E}\left\{ \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 \right\}\,.$$

A well known result in estimation theory [63] is that the MMSE or the *Bayes' estimator* is the conditional expectation of the random variable given the measurement, i.e.,

$$\hat{\mathbf{x}} = \mathrm{E}_{\mathbf{x}} \left\{ \mathbf{x} \mid \mathbf{y} = \mathbf{y} \right\} . \tag{2.4}$$

Writing out the expectation and using Bayes' rule results in

$$\begin{aligned} \hat{\mathbf{x}} &= \mathrm{E}_{\mathbf{x}} \left\{ \mathbf{x} \mid \mathbf{y} = \mathbf{y} \right\} \\ &= \int_{\mathbb{R}^N} \mathbf{x} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x} \mid \mathbf{y}) d\mathbf{x} \\ &= \frac{1}{f_{\mathbf{y}}(\mathbf{y})} \int_{\mathbb{R}^N} \mathbf{x} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} . \end{aligned}$$

Inserting the posterior probability function $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})$ that results from the measurement model (2.1) and the additive noise pdf (2.2) gives

$$\hat{\mathbf{x}} = \frac{1}{f_{\mathbf{y}}(\mathbf{y})} \int_{\mathbb{R}^N} \mathbf{x} \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_{\mathsf{w}}^2}} \exp\left(-\frac{(y_m - (\mathbf{A}\mathbf{x})_m)^2}{2\sigma_{\mathsf{w}}^2} \right) \prod_{n=1}^{N} f_{\mathsf{x}}(x_n) d\mathbf{x} . \tag{2.5}$$

Apart from some special cases (e.g., if $\mathbf{x}$ follows a multivariate Gaussian distribution or is a discrete random vector), this integral cannot be carried out component-wise for each n since the posterior pdf requires the full vector $\mathbf{x}$ for each component m. Therefore the integration is infeasible even when the problem dimension is only moderate. BAMP offers an approximate solution of the integral in (2.5) via loopy belief propagation [35, 64, 65].

2.1 The BAMP Algorithm

We first state the BAMP algorithm including initialization and stopping criterion in its most general form in Algorithm 1. Next, the quantities involved and their roles are discussed:

- $\mathbf{u}^{(t)}$, *decoupled measurement*: in [36] an intuitive interpretation was suggested named the *decoupling principle*. Even though the measurement $\mathbf{y}$ has (only) dimension M (typically $M < N$), the measurement model (2.1) is replaced by an equivalent *decoupled measurement model*

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \quad \Leftrightarrow \quad \mathbf{u}^{(t)} = \mathbf{x} + \mathbf{v}^{(t)} , \tag{2.6}$$

Algorithm 1 Bayesian approximate message passing

Input: $t = 0\,, \hat{\mathbf{x}}^{(t)} = \mathbf{0}_{N\times 1}\,, \mathbf{z}^{(t)} = \mathbf{y}$

do:

1: $t \leftarrow t+1$ ▷ increment iteration counter

2: $\sigma_{\mathsf{v}}^{2(t-1)} = \mathrm{Var}(z_m^{(t-1)})$ ▷ estimate effective noise variance

3: $\mathbf{u}^{(t-1)} = \hat{\mathbf{x}}^{(t-1)} + \mathbf{A}^T\mathbf{z}^{(t-1)}$ ▷ decouple measurements

4: $\forall n \in [N]$: $\hat{x}_n^{(t)} = F\left(u_n^{(t-1)}; \sigma_{\mathsf{v}}^{2(t-1)}\right)$ ▷ estimation

5: $\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}^{(t)} + \frac{1}{M}\mathbf{z}^{(t-1)} \sum_{n=1}^{N} F'\left(u_n^{(t-1)}; \sigma_{\mathsf{v}}^{2(t-1)}\right)$ ▷ calculate residual

while stopping criterion is false

Output: $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(t)}$

with $\mathbf{v}^{(t)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathsf{v}}^{2(t)}\mathbf{I}_N)$. Note that the measurement noise $\mathbf{w}$ in (2.1) has power σ_{w}^2. In the decoupled model, due to the interference that arises from the fact that $\mathbf{A}$ has mostly nonzero entries, $\sigma_{\mathsf{v}}^{2(t)} \geq \sigma_{\mathsf{w}}^2$. BAMP is designed such that x_n and $\mathsf{v}_n^{(t)}$ are independent. It follows that the pdf of $\mathsf{u}_n^{(t)}$ reads [66, Chapter 6.2]

$$f_{\mathsf{u}_n^{(t)}}(u_n^{(t)}) = f_{\mathsf{u}^{(t)}}(u_n^{(t)}) = f_{\mathsf{x}}(u_n^{(t)}) * \mathcal{N}(u_n^{(t)}; 0, \sigma_{\mathsf{v}}^{2(t)})\,,$$

i.e., the convolution of the signal prior pdf with a Gaussian pdf.

- $\sigma_{\mathsf{v}}^{2(t)}$, *effective noise variance*: the effective noise variance arises as the sum of the additive noise variance and the *interference* or *undersampling noise* that results from the mixing nature of the linear measurement. It is calculated as the mean empirical power of the residual $\mathbf{z}^{(t-1)}$.

- $\hat{\mathbf{x}}^{(t)}$, *current signal estimate*: the estimator function $F(\cdot;\cdot)$ is the MMSE estimator or *denoiser* that acts on the decoupled measurement $u_n^{(t)}$, with parameter $\sigma_{\mathsf{v}}^{2(t)}$. It is designed specifically for the signal prior pdf $f_{\mathsf{x}_n}(x_n)$ and additive Gaussian noise:

$$F(u_n^{(t)}; \sigma_{\mathsf{v}}^{2(t)}) = \mathrm{E}\left\{\mathsf{x}_n \mid \mathsf{u}_n^{(t)} = u_n^{(t)}\right\}\,. \tag{2.7}$$

- $\mathbf{z}^{(t)}$, *residual*: the residual is in essence the mismatch between the measurement $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$ and the measured version of the estimated signal $\mathbf{A}\hat{\mathbf{x}}^{(t)}$. The last term in Step 5 of Algorithm 1, i.e.,

$$\frac{1}{M}\mathbf{z}^{(t-1)} \sum_{n=1}^{N} F'\left(u_n^{(t-1)}; \sigma_{\mathsf{v}}^{2(t-1)}\right)\,,$$

is called the *Onsager term* and renders the effective noise $\mathsf{v}_n = \mathsf{u}_n - \mathsf{x}_n$ Gaussian distributed and guarantees the independence of x_n and v_n as $N \to \infty$ [36, 39].

- $\hat{\mathbf{x}}$, *output estimate*: in practice, since computational power and time is limited, a suitable stopping criterion is declared in order to terminate the algorithm and use the current estimate as the final estimate. Typically, one can use two conditions for termination (possibly simultaneously): a) the relative change in the value of a set of variables is small enough and its accuracy is judged sufficient, b) the prescribed maximum number of iterations is reached. Formally, for $t \geq 1$

$$\text{stop if} \quad \|\hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}^{(t-1)}\|_2^2 \leq \varepsilon_{\text{tol}} \|\hat{\mathbf{x}}^{(t-1)}\|_2^2 \quad \text{or} \quad t \geq t_{\max} \tag{2.8}$$

  with a small $\varepsilon_{\text{tol}} > 0$.

As the iterations proceed, $\hat{\mathbf{x}}^{(t)}$ approaches the MMSE estimate (2.4), which is in general not equal to the *maximum a posteriori* (MAP) estimate

$$\hat{\mathbf{x}} = \operatorname*{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^N} \mathrm{P}\{\tilde{\mathbf{x}} \mid \mathbf{y} = \mathbf{y}\} \ .$$

The difference between the MAP and the MMSE estimator (in our case BAMP) becomes apparent when the components of the unknown $\mathbf{x}$ take discrete values. Prominently in CS, a large fraction of the components is expected to be exactly zero. Due to BAMP being an MMSE estimator, the final estimate almost never hits a discrete value exactly. Thus, when discrete values are involved, post-processing is necessary: this can be as simple as nearest-neighbor search (quantization), or more complex such as the *expectation-maximization* (EM) algorithm [67], which is elaborated in Section 3.5.
Another noteworthy property of the BAMP algorithm is its low computational complexity. It does not involve matrix inversions. Once the estimator function $F(\cdot;\cdot)$ is available, the algorithm requires only matrix-vector multiplications and additions. This makes it not only an attractive choice from a theoretical viewpoint but also easily implementable in practice [68].

Zero-mean prior In general, the prior knowledge or assumption about the unknown is in form of a pdf $f_{\mathsf{x}}(x_n)$. If $\boldsymbol{\mu}_x = \mathrm{E}\{\mathbf{x}\} \neq \mathbf{0}$, i.e., the unknown has a nonzero mean,

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{w} \\ &= \mathbf{A}\boldsymbol{\mu}_x + \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x) + \mathbf{w} \,, \end{aligned}$$

which results in

$$\bar{\mathbf{y}} = \mathbf{y} - \mathbf{A}\boldsymbol{\mu}_x = \mathbf{A}\bar{\mathbf{x}} + \mathbf{w}\,, \tag{2.9}$$

where $\mathrm{E}\{\bar{\mathbf{x}}\} = \mathbf{0}$. That is, every measurement of a nonzero-mean random variable can be recast as an equivalent measurement of a transformed zero-mean random variable characterized by a pdf $f_{\mathsf{x}}(x_n - \mu_n)$ centered at 0. We highlight that this does not influence the performance or the convergence behavior of BAMP.
Also note that if the unknown $\mathbf{x}$ is expected to be zero-mean, the decoupled measurement $\mathbf{u}^{(t)}$, the residual vector $\mathbf{z}^{(t)}$, and the current estimate vector $\hat{\mathbf{x}}^{(t)}$ are zero-mean and preserve this property across iterations.

2.2 Priors of Interest

2.2.1 Bernoulli-Gauss Prior

The BG pdf is defined as

$$f_{\mathsf{x}}(x_n) = (1-\epsilon)\delta(x_n) + \epsilon\mathcal{N}(x_n; 0, \sigma_{\mathsf{x}}^2)\,,$$

with parameters *nonzero probability* ϵ and complementary *sparsity* (or *zero* probability) $1-\epsilon$, and variance of the nonzero components σ_{x}^2. In CS, typically, $0 < \epsilon \ll 1$. Note that the sparsity K is not deterministic anymore, but is controlled by the zero probability $(1-\epsilon)$. In practice, one can assume $K \approx (1-\epsilon)N$. An example for a compressible signal with BG prior is depicted in Figure 2.1. The BG prior is a very powerful tool to model compressible signals, because when only a small fraction of the entries is nonzero, the remaining nonzero components will mostly fit the Gaussian distribution (a mismatch analysis between the BG and the Bernoulli-Laplace prior can be found in [60]). Furthermore, the BAMP with BG signals allows for an elegant analysis via multiple tools and extensive analytical results, which are discussed in Chapter 3. The MMSE estimator (2.7) is (neglecting the iteration index t and the component index n)

$$F(u; \sigma_{\mathsf{v}}^2) = \mathrm{E}\left\{\mathsf{x} \mid \mathsf{u} = u\right\}\,,$$

where $\mathsf{u} = \mathsf{x} + \mathsf{v}$. Using

$$\mathrm{Var}\left\{\mathsf{x} \mid \mathsf{x} \neq 0\right\} = \sigma_{\mathsf{x}}^2\,, \quad \mathrm{Var}\left\{\mathsf{v}\right\} = \sigma_{\mathsf{v}}^2\,, \quad \text{and} \quad \mathrm{Var}\left\{\mathsf{u} \mid \mathsf{x} \neq 0\right\} = \sigma_{\mathsf{x}}^2 + \sigma_{\mathsf{v}}^2 = \sigma_{\mathsf{u}}^2\,,$$

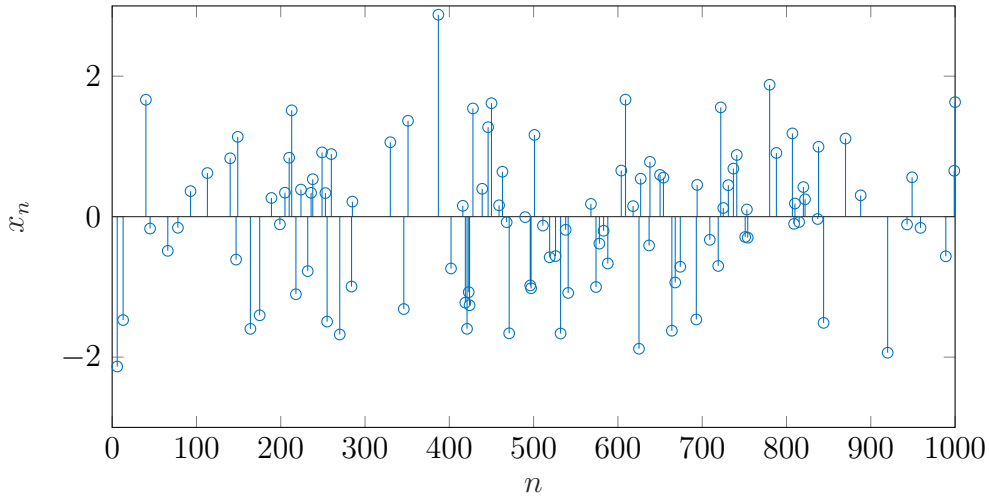


Fig. 2.1 A realization of a BG signal (dimension $N = 1000$, nonzero probability $\epsilon = 0.1$, variance $\sigma_{\mathsf{x}}^2 = 1$).

the final result reads

$$\begin{aligned} F(u;\sigma_{\mathsf{v}}^2) &= \frac{\epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)}{(1-\epsilon)\mathcal{N}(u;0,\sigma_{\mathsf{v}}^2)+\epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)}\sigma_{\mathsf{x}}^2\sigma_{\mathsf{u}}^{-2}u \\ &= \frac{F_N(u;\sigma_{\mathsf{v}}^2)}{F_D(u;\sigma_{\mathsf{v}}^2)}\sigma_{\mathsf{x}}^2\sigma_{\mathsf{u}}^{-2}u\,, \end{aligned}$$

with numerator and denominator

$$\begin{aligned} F_N(u;\sigma_{\mathsf{v}}^2) &= \epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)\,, \\ F_D(u;\sigma_{\mathsf{v}}^2) &= (1-\epsilon)\mathcal{N}(u;0,\sigma_{\mathsf{v}}^2)+\epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)\,. \end{aligned}$$

Its derivative is derived as

$$\begin{aligned} F'(u;\sigma_{\mathsf{v}}^2) &= \frac{d}{du}F(u;\sigma_{\mathsf{v}}^2) \\ &= \frac{1}{F_D(u;\sigma_{\mathsf{v}}^2)}\Bigg(\epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)\Big(\sigma_{\mathsf{x}}^2\sigma_{\mathsf{u}}^{-2}-\sigma_{\mathsf{x}}^2\sigma_{\mathsf{u}}^{-4}u^2\Big) \\ &\quad+\Big((1-\epsilon)\mathcal{N}(u;0,\sigma_{\mathsf{v}}^2)\sigma_{\mathsf{v}}^{-2}+\epsilon\mathcal{N}(u;0,\sigma_{\mathsf{u}}^2)\sigma_{\mathsf{u}}^{-2}\Big)F(u;\sigma_{\mathsf{v}}^2)u\Bigg)\,. \end{aligned}$$

An example of the estimator and its derivative for the BG signal prior is depicted in Figure 2.2. Observe that values close to zero, as they most probably correspond to zero components, are dampened strongly and set to almost zero. As the argument

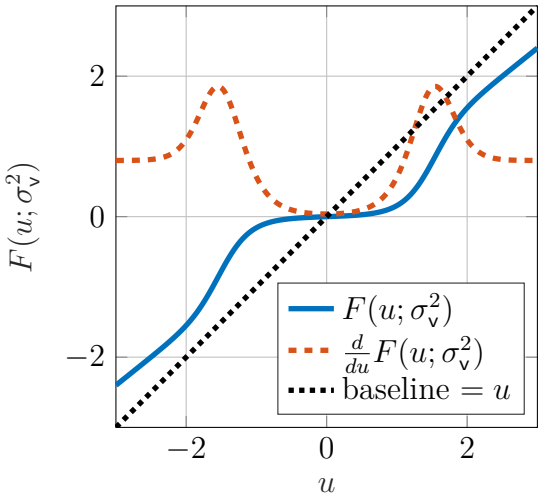


Fig. 2.2 MMSE estimator ($F(u;\sigma_{\mathsf{v}}^2)$) and its derivative for the BG prior (nonzero probability $\epsilon = 0.1$, variance $\sigma_{\mathsf{x}}^2 = 1$, noise variance $\sigma_{\mathsf{v}}^2 = 0.25$).

gets larger, distinguishing between activity and non-activity becomes harder and the dampening turns softer. When the argument is large even compared to the nonzero signal variance σ_{x}^2, the dampening gets stronger again in order to remove the additive noise from the decoupled measurement of the probably nonzero component.

2.2.2 Discrete Prior

The discrete prior pdf is defined as

$$f_{\mathsf{x}}(x_n) = \sum_{c=1}^{C} \epsilon^{(c)} \delta(x_n - s^{(c)}) \,,$$

where $\mathcal{S} = \{s^{(1)}, \ldots, s^{(C)}\}$ is the symbol alphabet of size $C = |\mathcal{S}|$, and $\epsilon^{(c)}$ ($c \in [C]$) are the individual symbol probabilities which sum up to 1:

$$P\left\{x_n = s^{(c)}\right\} = \epsilon^{(c)} \,, \quad \text{with} \quad \sum_{c=1}^{C} \epsilon^{(c)} = 1 \,.$$

In CS, typically, w.l.o.g. $s^{(1)} = 0$ and $1 > \epsilon^{(1)} \gg 0$. Note that if there is a dominant symbol that is nonzero, the measurement can be transformed into an equivalent measurement following the procedure similar to the mean removal in Section 2.1 in order to obtain a CS measurement with 0 as the dominant symbol. This, however, does not influence the performance of BAMP. An example discrete prior signal is depicted in Figure 2.3. The discrete prior is a powerful tool for, e.g., telecommunication

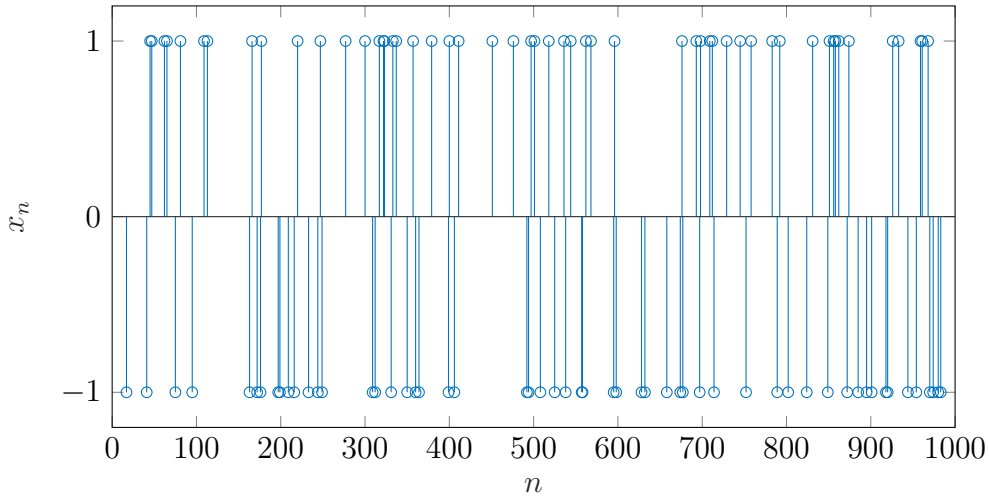


Fig. 2.3 A realization of a signal with discrete prior (dimension $N = 1000$, $\mathcal{S} = \{0, -1, 1\}$, $\epsilon^{(1)} = 0.9\,, \epsilon^{(2)} = \epsilon^{(3)} = 0.05$).

applications, where the set of possible transmit symbols constitute a finite *symbol alphabet* [69, 70]. The MMSE estimator function for the discrete prior pdf reads

$$F(u; \sigma_{\mathsf{v}}^2) = \frac{\sum_{c=1}^{C} \epsilon^{(c)} s^{(c)} \mathcal{N}(u; s^{(c)}, \sigma_{\mathsf{v}}^2)}{\sum_{c=1}^{C} \epsilon^{(c)} \mathcal{N}(u; s^{(c)}, \sigma_{\mathsf{v}}^2)}\,,$$

which is essentially a weighted sum of all symbols. Its derivative calculates as

$$\begin{aligned} F'(u; \sigma_{\mathsf{v}}^2) &= \frac{d}{du} F(u; \sigma_{\mathsf{v}}^2) \\ &= \sigma_{\mathsf{v}}^{-2} \left(\frac{\sum_{c=1}^{C} \epsilon^{(c)} {s^{(c)}}^2 \mathcal{N}(u; s^{(c)}, \sigma_{\mathsf{v}}^2)}{\sum_{c=1}^{C} \epsilon^{(c)} \mathcal{N}(u; s^{(c)}, \sigma_{\mathsf{v}}^2)} - F(u; \sigma_{\mathsf{v}}^2) \right)\,. \end{aligned}$$

An example of the estimator and its derivative for the discrete signal prior is depicted in Figure 2.4.

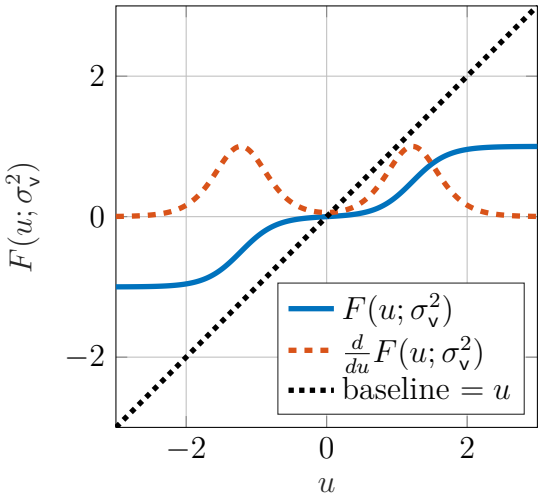


Fig. 2.4 MMSE estimator ($F(u;\sigma_{\mathsf{v}}^2)$) and its derivative for the discrete prior ($\mathcal{S} = \{0,-1,1\}$, $\epsilon^{(1)} = 0.9$, $\epsilon^{(2)} = \epsilon^{(3)} = 0.05$, noise variance $\sigma_{\mathsf{v}}^2 = 0.25$).

2.3 Analysis

2.3.1 State Evolution

A very advantageous property of BAMP is that its performance can be analytically predicted. The corresponding framework is termed SE [39, 71]. In particular, the SE equations are capable of describing the evolution of the expected effective noise variance (the *state*), and thus the expected MSE, over iterations. The BAMP algorithm is equivalent to a dynamical system described by the SE equation with state $\sigma_{\mathsf{v}}^{2(t)}$. The SE prediction becomes exact as $M, N \to \infty$ as the sampling rate $R = \frac{M}{N}$ is constant. In practice, however, its results are sufficiently accurate even for moderately large dimensions (e.g., $N = 1000$). This can be used to, e.g.,

- predict the average performance of BAMP given a set of parameters (nonzero probability ϵ, sampling rate $R = \frac{M}{N}$, noise level σ_{w}^2, and signal prior $f_{\mathsf{x}}(x_n)$),
- find the required sampling rate $R = \frac{M}{N}$ in order to achieve a desired performance.

For any signal prior $f_{\mathsf{x}}(x_n)$ and any fixed (not necessarily the MMSE) estimator function $F(u_n;\sigma_{\mathsf{v}}^2)$, additive noise variance σ_{w}^2 and effective noise variance $\sigma_{\mathsf{v}}^{2(t)}$, the input-output relationship between the successive effective noise variances (states) is

$$\sigma_{\mathsf{v}}^{2(t+1)} = S(\sigma_{\mathsf{v}}^{2(t)}) = \sigma_{\mathsf{w}}^2 + \frac{1}{R}\underbrace{\mathrm{E}_{\mathsf{x},\mathsf{v}}\left\{\left(F(\mathsf{x}+\mathsf{v};\sigma_{\mathsf{v}}^{2(t)}) - \mathsf{x}\right)^2\right\}}_{\widehat{\mathrm{MSE}}(\hat{\mathbf{x}}^{(t)},\mathbf{x})}, \tag{2.10}$$

where $\mathsf{v} \sim \mathcal{N}(0, \sigma_{\mathsf{v}}^{2(t)})$. The initial estimate reads

$$\sigma_{\mathsf{v}}^{2(0)} = \sigma_{\mathsf{w}}^2 + \frac{1}{R} \mathrm{E}_{\mathsf{x}} \left\{ \mathsf{x}^2 \right\} = \mathrm{Var}\{\mathsf{y}_m\} \,.$$

Note that the BAMP algorithm delivers an estimate of the corresponding MSE for every signal estimate $\hat{\mathbf{x}}^{(t)}$ with $\sigma_{\mathsf{v}}^{2(t)}$:

$$\widehat{\mathrm{MSE}}(\hat{\mathbf{x}}^{(t)}, \mathbf{x}) = \mathrm{E}_{\mathsf{x},\mathsf{v}} \left\{ \left(F(\mathsf{x} + \mathsf{v}; \sigma_{\mathsf{v}}^{2(t)}) - \mathsf{x} \right)^2 \right\} \,.$$

This feature of the BAMP algorithm has been exploited in a number of works, e.g., for joint sparsity [47, 48] and reconstruction without prior information [72]. In general, computing the integral involved in (2.10) is infeasible. Nonetheless, solving it numerically and visualizing the results is crucial in understanding the behavior of BAMP. Details on the numerical evaluation can be found in Appendix D.

Discussion

In order to understand the behavior of BAMP, one relies on the fact that it can be interpreted as a dynamical system whose state at (discrete) time t is $\sigma_{\mathsf{v}}^{2(t)}$. The stationary (fixed) points of the SE equation (2.10) correspond the stationary points of BAMP, i.e., where $\sigma_{\mathsf{v}}^{2(t+1)} = \sigma_{\mathsf{v}}^{2(t)}$ and $\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)}$. Roughly speaking, a stable fixed point is a fixed point to which the system converges from an arbitrarily small neighborhood of that point; whereas an unstable fixed point is a fixed point to which the system does not converge from an arbitrarily small neighborhood of that point. In Figure 2.5 the (continuous) SE curves are depicted for different sampling rates R and the BG prior with nonzero probability $\epsilon = 0.1$. The observer can distinguish between two rate regions:

1. High rate region, e.g., $R = 0.4$ (Figure 2.5a): the SE curve stays below the baseline and they have a single intersection, i.e., one stable fixed point, in $\sigma_{\mathsf{v}}^{2(t+1)} = \sigma_{\mathsf{v}}^{2(t)} = \sigma_{\mathsf{w}}^2 = 0$. (Note that this is not visible on the double logarithmic plot, but as the SE curve stays below the baseline, one can be sure that $(0, 0)$ is a stable fixed point.) That is, in expectation BAMP converges to $\sigma_{\mathsf{v}}^{2(t)} = \sigma_{\mathsf{w}}^2 = 0$ and $\hat{\mathbf{x}}^{(t)} \to \mathbf{x}$, i.e., $\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}, \mathbf{x}) \to 0$.

2. Low rate region, e.g., $R \lesssim 0.21$ (Figure 2.5b-d): SE has a stable fixed point at $\sigma_{\mathsf{v}}^{2(t)} = 0$ and a second stable fixed point at high effective noise (e.g., for $R = 0.165$: $\sigma_{\mathsf{v}}^{2(t)} \approx -5\mathrm{dB}$). The two stable fixed points are separated by an

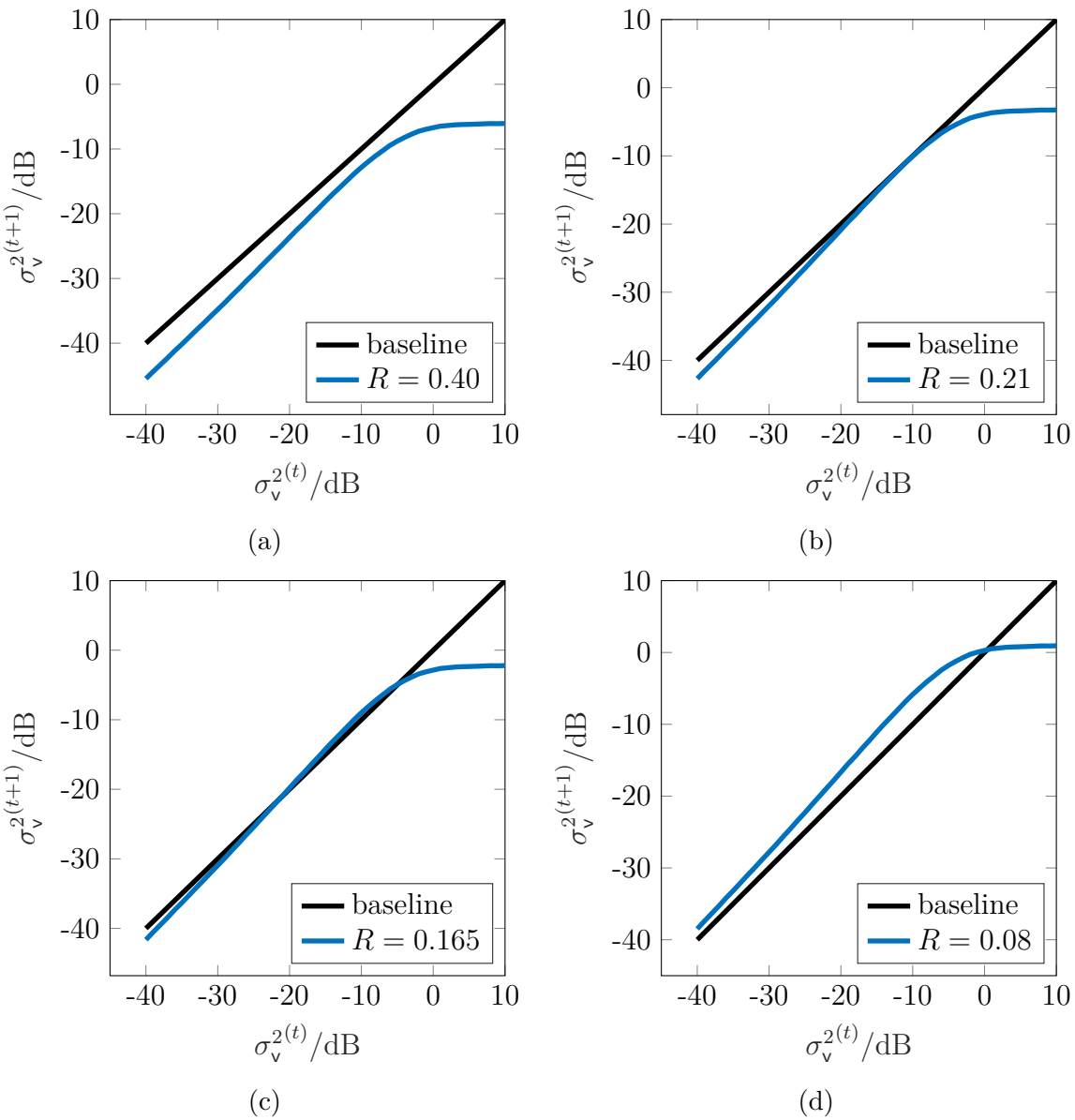


Fig. 2.5 Empirical noiseless SE curves for the BG prior and estimator obtained from Monte Carlo simulation (dimension $N = 1000$, $\epsilon = 0.1$, $\sigma_{\mathsf{w}}^2 = 0$).

unstable fixed point. BAMP typically reaches the fixed point with the higher effective noise variance. Note that at $R = 0.08$ V-BAMP reached the PT of *second order*, where only two fixed points exist, both stable: one at $\sigma_{\mathsf{v}}^{2(t)} = 0$ and one at a relatively high effective noise variance (in this particular case $\sigma_{\mathsf{v}}^2 \approx 0$dB).

The rate that separates the high and low rate regions according to the above classification is called the PT rate and is denoted by R_{PT}. It separates $\mathbb{R}^+$ into two regions: for $R < R_{\mathrm{PT}}$, BAMP reaches a relatively high $\sigma_{\mathsf{v}}^{2(t)}$ (and thus MSE) and

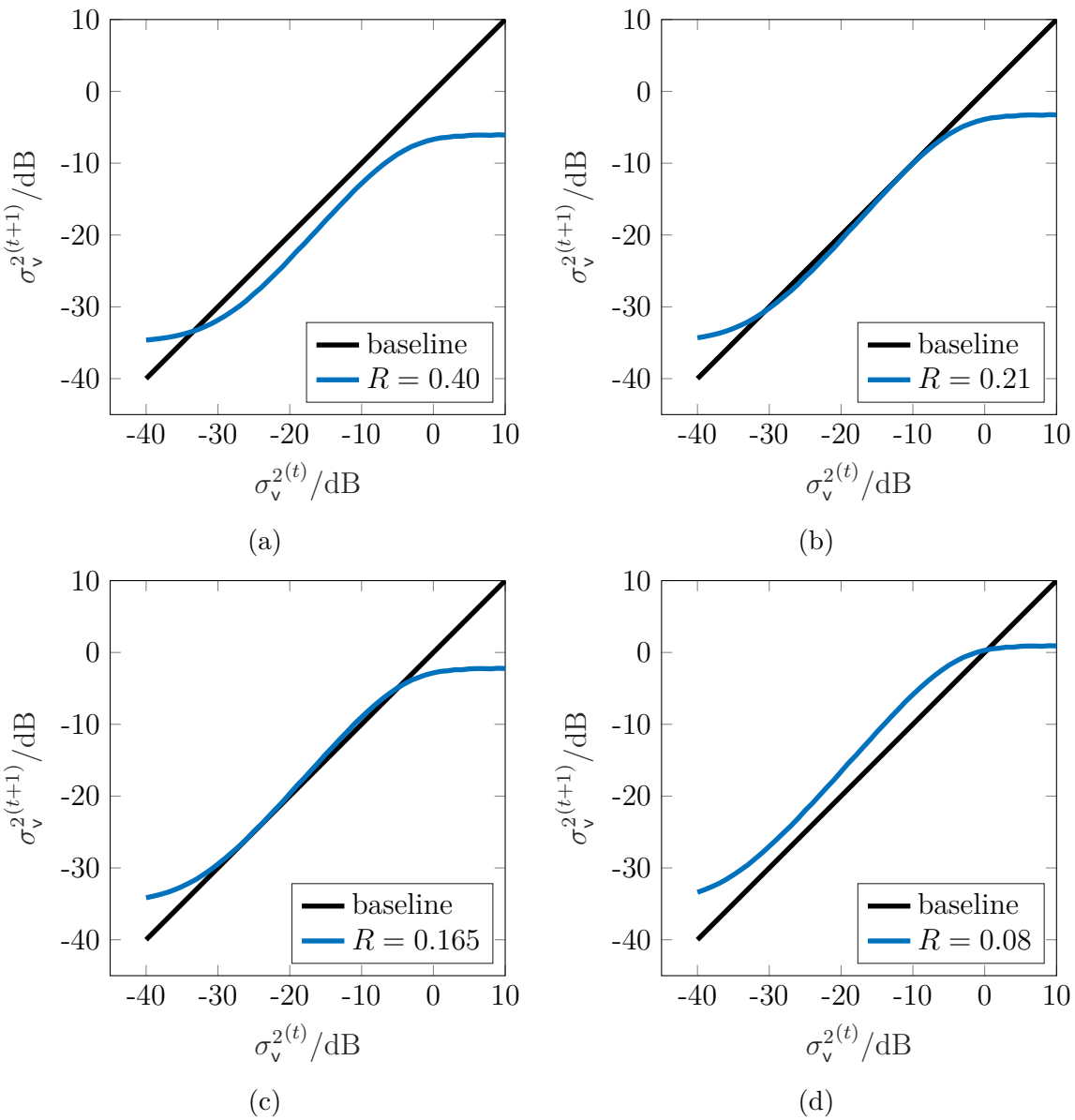


Fig. 2.6 Empirical noisy SE curves for the BG prior and estimator obtained from Monte Carlo simulation (dimension $N = 1000$, $\epsilon = 0.1$, $\sigma_{\mathsf{w}}^2 = -35$dB).

we call the recovery *unsuccessful*. for $R > R_{\mathrm{PT}}$, BAMP reaches a relatively low $\sigma_{\mathsf{v}}^{2(t)}$ (and thus MSE) and we call the recovery *successful*. In Figure 2.6 additive noise with $\sigma_{\mathsf{w}}^2 = -35$dB was added, and analogously to the noiseless case, the PT does occur with a similar R_{PT}.

In Figure 2.7 SE curves for the BG signal prior are shown, for $R = 0.21 \approx R_{\mathrm{PT}}$ and different noise levels. Observe that as the additive noise variance increases, the SE curve flattens, i.e., it shifts away from having a nearly parallel section with the baseline.

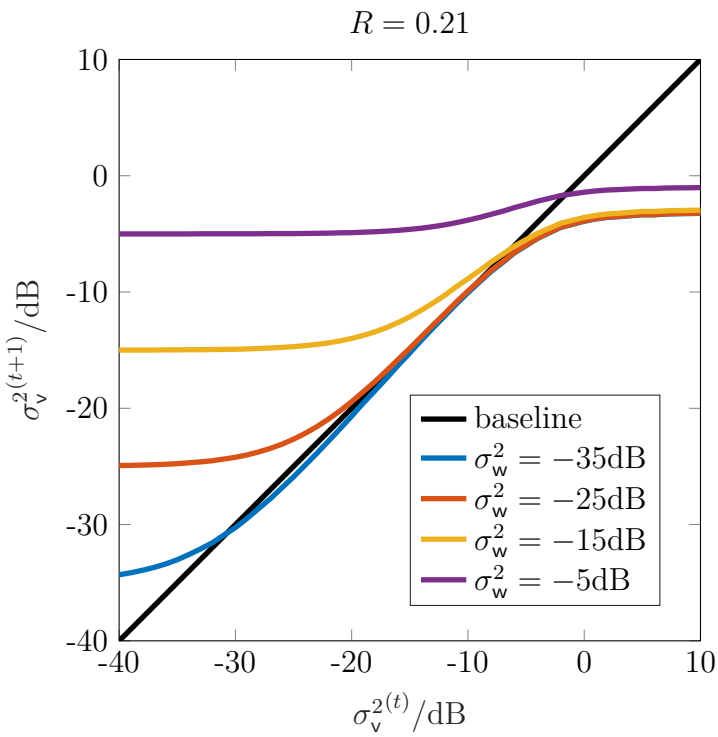


Fig. 2.7 Empirical noisy SE curves for the BG prior and estimator with rate $R = 0.21$ and different additive noise levels.

That is, it does not have two crossings with the baseline and the PT does not occur. We conclude that in the noiseless case (and the very low noise level case) BAMP is characterized solely by the PT rate R_{PT}: at $R > R_{\mathrm{PT}}$ BAMP is successful with high probability, while at $R < R_{\mathrm{PT}}$ it is unsuccessful with high probability. Furthermore, with increasing additive noise, the PT (i.e., the sudden drop in the MSE with increasing rate R) vanishes.

2.3.2 Phase Transition Curves

In the case of noiseless CS measurement

$$\mathbf{y} = \mathbf{A}\mathbf{x}\,,$$

the average evolution of BAMP is characterized by the sampling rate R and the prior pdf $f_{\mathsf{x}}(x_n)$. When the prior pdf models a sparse signal (as in the case of the BG pdf with a high zero probability $1-\epsilon$) and can be characterized by the nonzero probability ϵ, BAMP is analyzed as a function of the pair (R, ϵ). Typically, for a pair (R, ϵ) BAMP either does or does not converge to the correct solution with probability close

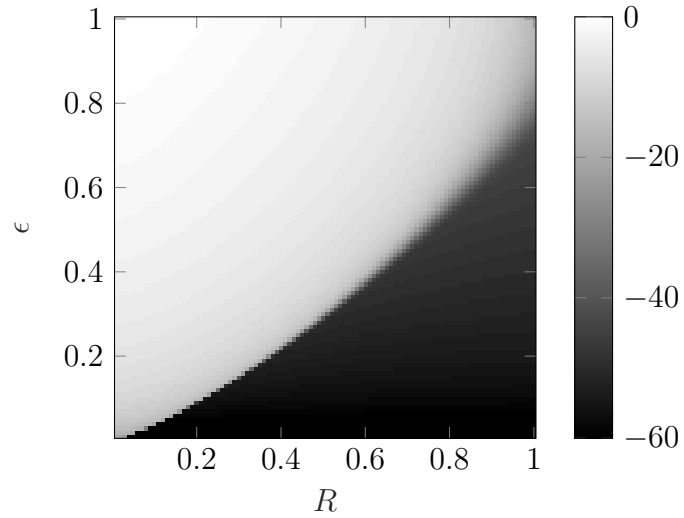


Fig. 2.8 MSE of BAMP as predicted by SE, sampling rate R versus nonzero probability ϵ with BG signal prior. The darkness of the shades corresponds to the $\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}, \mathbf{x})/\mathrm{dB}$ as $t \to \infty$. The PT curve is what separates the dark region from the bright region in the low ϵ regime.

to one. For a given nonzero probability ϵ, BAMP with $R > R_{\mathrm{PT}}$ will converge to the correct solution whp, whereas for $R < R_{\mathrm{PT}}$ BAMP will fail to find the correct solution whp. The two regions in the sampling rate - sparsity plane are separated by the *phase transition curve* (PTC) [39, 73, 74]. In Figure 2.8 the MSE as a function of the sampling rate R and the nonzero probability ϵ is depicted for a BG signal and correspondingly designed BAMP. The darkness of the shades corresponds to $\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}})/\mathrm{dB}$. The PTC can be easily identified as the line separating the dark and the bright region in the low ϵ regime: with parameters in the dark region, BAMP converges to $\mathbf{x}$ whp, whereas for parameters in the bright region, BAMP fails to converge to $\mathbf{x}$ whp. Observe that the transition from the low MSE to high MSE region is smoother as the the nonzero probability ϵ grows out of the CS region, i.e., the PT description is valid only in the CS regime, where $\epsilon \ll 1$.

Note that above the SE and PT analysis are precisely valid only for the signals with BG prior and its corresponding MMSE estimator. The SE curves for other signals and other estimators might look differently. However, in the CS regime, where the zero probability is high, the deviation in terms of the fixed points of the SE equation and of the PT curves is expected to be minor.

Chapter 3

Vector Bayesian Approximate Message Passing

A practically relevant and theoretically challenging extension of the original CS problem arises when multiple measurements of the form

$$\mathbf{y}(b) = \mathbf{A}(b)\mathbf{x}(b) + \mathbf{w}(b)\,, \quad b \in [B] \tag{3.1}$$

are obtained. Here, the unknowns have the same dimension $\mathbf{x}(b) \in \mathbb{R}^N$. When the B measured vectors $\mathbf{x}(b)$, $b \in [B]$, are not completely independent, it is a promising path to perform joint recovery, i.e., incorporate the dependencies between the different measurements. A prominent case of dependency is subsumed under the *joint sparsity models* (JSMs) [49], which require that the nonzero patterns of the b measured vectors are overlapping in some sense. Suppose that after measurement one performs B independent recovery procedures to obtain B estimates. In the low noise and ideal sparsity-sampling rate regime the result is typically satisfying, but with increasing noise or decreasing sampling rate all individual recoveries will fail to succeed. It was proved to be advantageous, however, to modify existing recovery methods such that they exploit the knowledge of the mutual support, which results in significant broadening of the supported *signal-to-noise ratio* (SNR) and sampling rate regime.

In the probabilistic setting, the measured vectors $\mathbf{x}(b)$ are realizations of a random vector $\mathsf{x}(b)$. Let us denote the column vector constituted by the identically indexed components of the B measured vectors by $\vec{\mathbf{x}}_n = (x_n(1), \ldots, x_n(B))^T$. In order to incorporate the dependencies between the measured vectors, a joint pdf similar to (2.3)

for the random vector $\vec{\mathbf{x}}_n$ is introduced, which in general satisfies

$$f_{\vec{\mathbf{x}}_n}(\vec{\mathbf{x}}_n) \neq \prod_{b=1}^{B} f_{\mathsf{x}_{n,b}}(x_n(b))\,.$$

The concept of joint sparsity in CS can be incorporated by the multivariate Dirac delta in the prior pdf, i.e., it is assumed that

$$f_{\vec{\mathbf{x}}_n}(\vec{\mathbf{x}}_n) = f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = (1-\epsilon)\delta(\vec{\mathbf{x}}_n) + \epsilon f_{nz}(\vec{\mathbf{x}}_n)$$

i.i.d. over the n components and with $0 < \epsilon \ll 1$. Here, $f_{nz} \neq \delta$ is the distribution of $\vec{\mathbf{x}}_n$ given that $n \in \mathrm{supp}(\mathbf{x}(b))$, $b \in [B]$.

3.1 Motivation and Overview

Joint sparsity arises in a number of applications: multiuser communications [62], RFID [60], and (medical) imaging [55–58]. The notion of joint sparsity can be precisely specified by the JSMs [49]:

JSM-1: sparse common support with sparse innovations. JSM-1 admits the following representation for the set of measured signals:

$$\mathbf{x}(b) = \mathbf{x}_{\mathrm{c}} + \mathbf{x}_{\mathrm{i}}(b)\,, \quad \forall b\,.$$

Here, $\mathbf{x}_{\mathrm{c}}$ is common for all B signal vectors and sparse, and $\mathbf{x}_{\mathrm{i}}(b)$ is the innovation of each signal vector: in general they differ for the B signal vectors and are sparse. This allows us to model, e.g., a sensor network whose nodes measure a quantity for which a sparse representation is known. The common component is due to high temporal and geographical correlation, whereas the innovation components model temporally or spatially local effects. The innovation term can also represent sensor failures by, e.g., canceling nonzero components of the common component.

JSM-2: strictly common sparse support. JSM-2 admits the following representation for the set of measured signals:

$$\mathrm{supp}(\mathbf{x}(b)) = \mathcal{S} \quad \forall b\,, \tag{3.2}$$

where $\mathcal{S}$ is the common support of all B sparse vectors. This can model, e.g., a communication system in which a signal (sparse in time, frequency, or code domain) transmitted by a single node is captured by multiple separate receiver units (antennas).

JSM-3: nonsparse common support with sparse innovations. JSM-3 admits the following representation for the set of measured signals:

$$\mathbf{x}(b) = \mathbf{x}_{\mathrm{c}} + \mathbf{x}_{\mathrm{i}}(b)\,, \quad \forall b\,, \tag{3.3}$$

where $\mathbf{x}_{\mathrm{c}}$ is common for all signal vectors and is nonsparse, and $\mathbf{x}_{\mathrm{i}}(b)$ is the innovation of each signal vector: in general they differ for the B signal vectors and are sparse. This model can be useful, e.g., in case of a sensor network whose nodes aim to detect different sources while receiving a strong background signal. In video encoding, the difference between subsequent frames can be interpreted as sparse innovation, whereas the images themselves (or the average image) constitute a dense signal.

The focus of this chapter is JSM-2. A collection of greedy methods (see Section 1.2.2) generalized to solve the joint sparse recovery is introduced in [49, 75]. Convex optimization based approaches such as group LASSO are presented in [76, 77]. Methods that support higher problem dimensions are based on the probabilistic measurement model and in particular message passing. The authors of [47, 48] introduce a binary latent variable for each component n that indicates whether component n is nonzero or not, and perform BAMP once on each measurement. Each BAMP instance in each iteration delivers a likelihood on the underlying latent variable. The B likelihoods on latent variable n are then equalized and fed back into the B BAMP instances. This is a powerful yet simple method (also known as *turbo reconstruction*) that interconnects individual recoveries during their iterations, which leads to a very fast converging method and a low probability of disagreement over the nonzero patterns.
Generalizing the BAMP algorithm for JSM-2 delivers an efficient and fast approximate MMSE estimator algorithm. Moreover, it turns out to be applicable even to JSM-1 for certain priors, and also for general priors that do not necessarily reflect sparsity. Furthermore, it allows for arbitrary correlations between the signal vectors $\mathbf{x}(b)$ and the noise vectors $\mathbf{w}(b)$. The remaining part of this chapter is devoted to investigate the extension of the BAMP algorithm to the JSM-2.

3.2 MMV and DCS

A cardinal distinction can be made based on the construction of the measurement matrices $\mathbf{A}(b)$ in (3.1).

Distributed compressed sensing (DCS). In the DCS model, the measurement matrices are independent realizations of the same distribution, and thus differ in general, i.e., $\mathbf{A}(b) \neq \mathbf{A}(b')$, $\forall b' \neq b$. Furthermore, even the dimensions of each measurement can differ, i.e., $\mathbf{A}(1) \in \mathbb{R}^{M_1 \times N}, \ldots, \mathbf{A}(B) \in \mathbb{R}^{M_B \times N}$ with possibly different $M_1, \ldots, M_B$.

Multiple measurement vectors (MMV). In the MMV model, the measurement matrices coincide, i.e., $\mathbf{A}(b) = \mathbf{A}$, $\forall b \in [B]$. MMV can be also interpreted as measuring B-dimensional symbols instead of scalars. This simplification allows us to write (3.1) compactly as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}\,, \tag{3.4}$$

with

$$\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(B)) = \begin{pmatrix} \vec{\mathbf{y}}_1^T \\ \vdots \\ \vec{\mathbf{y}}_M^T \end{pmatrix} \in \mathbb{R}^{M \times B}\,,$$

$$\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(B)) = \begin{pmatrix} \vec{\mathbf{x}}_1^T \\ \vdots \\ \vec{\mathbf{x}}_N^T \end{pmatrix} \in \mathbb{R}^{N \times B}\,,$$

$$\mathbf{W} = (\mathbf{w}(1), \ldots, \mathbf{w}(B)) = \begin{pmatrix} \vec{\mathbf{w}}_1^T \\ \vdots \\ \vec{\mathbf{w}}_M^T \end{pmatrix} \in \mathbb{R}^{M \times B}\,.$$

Let us briefly discuss the differences in the recovery of jointly sparse vectors in the DCS and MMV scenarios via the following intuitive special cases:

1. Noiseless measurements of identical vectors:

$$\mathbf{x}(b) = \mathbf{x}\,, \quad \mathbf{w}(b) = \mathbf{0} \Leftrightarrow \begin{cases} \mathbf{y}(b) = \mathbf{A}(b)\mathbf{x} & \text{DCS}\,, \\ \mathbf{y}(b) = \mathbf{A}\mathbf{x} & \text{MMV}\,. \end{cases}$$

   Clearly, in the MMV case all measurements are identical, i.e., $\mathbf{y}(b) = \mathbf{y}(1)$

($b \in [B]$), and additional measurements are uninformative for the recovery. In the DCS case, however, one can stack both the measurement vectors $\mathbf{y}(b)$ and the matrices $\mathbf{A}(b)$ into a super-measurement

$$\begin{pmatrix} \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(B) \end{pmatrix} = \begin{pmatrix} \mathbf{A}(1) \\ \vdots \\ \mathbf{A}(B) \end{pmatrix} \mathbf{x}\,,$$

and obtain a measurement with a valid measurement matrix of dimension $BM \times N$ (note that the normalization of the stacked measurement matrix changes). For $B > 1$, this is an obvious advantage of DCS.

2. Noisy measurement of identical vectors:

$$\mathbf{x}(b) = \mathbf{x}\,, \quad \mathbf{w}(b) \neq \mathbf{0} \Leftrightarrow \begin{cases} \mathbf{y}(b) = \mathbf{A}(b)\mathbf{x} + \mathbf{w}(b) & \text{DCS}\,, \\ \mathbf{y}(b) = \mathbf{A}\mathbf{x} + \mathbf{w}(b) & \text{MMV}\,. \end{cases}$$

In the MMV case $\mathbf{y}(b)$ is the noisy observation of the same quantity $\mathbf{A}\mathbf{x}$. Thus, the possible advantage of having B measurements is the reduction of the additive noise via the averaging effect. In the DCS case, after stacking, one hase BM noisy measurements, and the noise reduction will be inherent to the recovery method since the sampling rate R is larger.

The most general case is noisy measurement of jointly sparse vectors, i.e., (3.1) in case of DCS and (3.1) with $\mathbf{A}(b) = \mathbf{A}$ ($b \in [B]$) in case of MMV. Intuitively, DCS is expected to have an advantage in terms of recovery accuracy over MMV because of the randomness it introduces in the different measurements: since all measured vectors $\mathbf{x}(b)$ select the same set of columns in the measurement matrices $\mathbf{A}$ respectively $\mathbf{A}(b)$, differently drawn measurement matrices can eliminate errors that arise due to, e.g., *too similar columns* selected by $\operatorname{supp}(\mathbf{x}(b))$ when the sampling rate is low. A counterargument that speaks for the better recovery performance of MMV is that in contrast to the mixing nature of DCS, MMV preserves correlation between the jointly measured vectors which can then be exploited.

Algorithm 2 Vector Bayesian approximate message passing for JSM-2

Input: $t = 0$, $\forall b \in [B]$: $\hat{\mathbf{x}}^{(t)}(b) = \mathbf{0}_{N\times 1}$, $\mathbf{z}^{(t)}(b) = \mathbf{y}(b)$

do:

1: $t \leftarrow t + 1$ ▷ increment iteration counter

2: $\Sigma_{\vec{\mathbf{v}}}{}^{(t-1)} = \begin{cases} \mathrm{Cov}(\vec{\mathbf{z}}_m^{(t-1)}) & \text{for MMV} \\ \mathrm{diag}(\mathrm{Cov}(\vec{\mathbf{z}}_m^{(t-1)})) & \text{for DCS} \end{cases}$ ▷ estimate effective noise covariance

3: $\forall b \in [B]$: $\mathbf{u}^{(t-1)}(b) = \hat{\mathbf{x}}^{(t-1)}(b) + \mathbf{A}^T(b)\mathbf{z}^{(t-1)}(b)$ ▷ decouple measurements

4: $\forall n \in [N]$: $\hat{\vec{\mathbf{x}}}_n^{(t)} = F\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t-1)}\right)$ ▷ estimation

5: $\forall m \in [M]$: $\vec{\mathbf{z}}_m^{(t)} = \vec{\mathbf{y}}_m - \left(\mathbf{A}(1)\hat{\vec{\mathbf{x}}}^{(t)}(1), \ldots, \mathbf{A}(B)\hat{\vec{\mathbf{x}}}^{(t)}(B)\right)_m$
$+ \frac{1}{M}\sum_{n=1}^{N} F'\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t-1)}\right)\vec{\mathbf{z}}_m^{(t-1)}$ ▷ calculate residual

while stopping criterion is false

Output: $\hat{\mathbf{x}}(b) = \hat{\mathbf{x}}^{(t)}(b)$, $\forall b \in [B]$

3.3 Vector BAMP for JSM-2

Just as the BAMP algorithm, V-BAMP acts on the probabilistic measurement model for vector-valued measurements:

$$\mathbf{y}(b) = \mathbf{A}(b)\mathbf{x}(b) + \mathbf{w}(b)\,.$$

The vectors $\vec{\mathbf{x}}_n$ follow a prior pdf i.i.d. over n:

$$f_{\vec{\mathbf{x}}_n}(\vec{\mathbf{x}}_n) = f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n)\,.$$

The noise $\vec{\mathbf{w}}_m$ is assumed to be zero-mean Gaussian with covariance $\Sigma_{\vec{\mathbf{w}}}$, i.e.,

$$\vec{\mathbf{w}}_m \sim \mathcal{N}(\mathbf{0}, \Sigma_{\vec{\mathbf{w}}})\,.$$

When the B measurement matrices are identical (MMV), one can write the measurements (3.1) compactly as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}\,.$$

From this form it is clear that the MMV model can be interpreted as mapping the set of N B-dimensional symbols $\vec{\mathbf{x}}_n$ onto the M B-dimensional symbols $\vec{\mathbf{y}}_m$. This fact is helpful in understanding the differences between V-BAMP for DCS and MMV, as the MMV case is a straightforward generalization of BAMP. We first state the V-BAMP algorithm [78] including initialization and stopping criterion in its most general form

in Algorithm 2. Next, the involved quantities and their roles are discussed, with focus on the differences to DCS:

- $\vec{\mathbf{u}}_n^{(t)}$, *decoupled vector measurements*: following the decoupling principle introduced in [36] and discussed in Section 2.1, the B individual decoupled measurements (2.6) can be extended to the vector/multivariate version directly in vector form as
$$\vec{\mathbf{u}}_n^{(t)} = \vec{\mathbf{x}}_n + \vec{\mathbf{v}}_n^{(t)} \quad \text{with} \quad \vec{\mathbf{v}}_n^{(t)} \sim \mathcal{N}(0, \Sigma_{\vec{\mathbf{v}}}) \,.$$

- $\Sigma_{\vec{\mathbf{v}}}^{(t)}$, *effective noise covariance*: in the MMV case, due to the fact that all B values $x_n(1), \ldots, x_n(B)$ are measured through the same matrix, their statistical properties carry over to $\vec{\mathbf{u}}_n$ (and also $\vec{\mathbf{y}}_m$ and $\vec{\mathbf{z}}_m$). Thus, V-BAMP considers the full covariance (second order statistics) of the residual, which is designed such that $\vec{\mathbf{u}}_n - \vec{\mathbf{x}}_n$ is zero-mean Gaussian with covariance matrix $\Sigma_{\vec{\mathbf{v}}}^{(t)}$. In the DCS case, however, due to the independence of the measurement matrices, correlations between signal components are eliminated, and the effective noise covariance consists of only the B individual variances.

- $\hat{\vec{\mathbf{x}}}_n^{(t)}$, *current signal estimate*: the estimator function $F(\cdot;\cdot)$ is the MMSE estimator that acts on the decoupled vector measurement $\vec{\mathbf{u}}_n^{(t)}$ with parameter $\Sigma_{\vec{\mathbf{v}}}^{(t)}$. It is designed specifically for the signal prior pdf $f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n)$ and additive Gaussian noise:
$$F(\vec{\mathbf{u}}_n^{(t)}; \Sigma_{\vec{\mathbf{v}}}^{(t)}) = \mathrm{E}\left\{\vec{\mathbf{x}}_n \mid \vec{\mathbf{u}}_n^{(t)} = \vec{\mathbf{u}}_n^{(t)}, \Sigma_{\vec{\mathbf{v}}}^{(t)}\right\} \,. \tag{3.5}$$

- $\mathbf{z}^{(t)}(b)$, *residual*: the residual vectors represent the mismatch between the measurement, i.e., $\mathbf{y}(b)$ and $\mathbf{A}(b)\hat{\mathbf{x}}^{(t)}(b)$. The last term in Step 5 of Algorithm 2, i.e.,
$$\frac{1}{M}\sum_{n=1}^{N} F'\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t-1)}\right) \vec{\mathbf{z}}_m^{(t-1)} \,,$$
the (matrix-valued) *Onsager term*, modifies the residuals $\vec{\mathbf{z}}_m^{(t)}$ such that the effective noise $\vec{\mathbf{v}}_n^{(t)} = \vec{\mathbf{u}}_n^{(t)} - \vec{\mathbf{x}}_n$ is zero-mean Gaussian with covariance $\Sigma_{\vec{\mathbf{v}}}^{(t)}$, and $\vec{\mathbf{v}}_n^{(t)}$ and $\vec{\mathbf{x}}_n$ independent, as $N \to \infty$.

- $\hat{\mathbf{x}}(b)$, *output estimate*: in practice, computational power and time is limited, and a suitable stopping criterion is used in order to terminate the algorithm and use the current estimate as the final estimate. The stopping criterion (2.8) can be

extended to, e.g.,

$$\text{stop if} \quad \sum_{b=1}^{B} \|\hat{\mathbf{x}}^{(t)}(b) - \hat{\mathbf{x}}^{(t-1)}(b)\|_2^2 \leq \varepsilon_{\text{tol}} \sum_{b=1}^{B} \|\hat{\mathbf{x}}^{(t-1)}(b)\|_2^2 \quad \text{or} \quad t \geq t_{\max}$$

with a small $\epsilon_{\text{tol}} > 0$.

Heterogeneous DCS

By assumption the jointly sparse vectors have dimension N, i.e., $\mathbf{x}(b) \in \mathbb{R}^N$, and the B measurement and noise vectors have the same dimension, i.e., $\mathbf{y}(b), \mathbf{w}(b) \in \mathbb{R}^M$, $b \in [B]$. When the B additive noise vectors are i.i.d., it is possible to extend the V-BAMP algorithm to the case when the measurement dimensions are not identical. That is, $\mathbf{y}(b), \mathbf{w}(b) \in \mathbb{R}^{M_b}$, with possibly different dimensions M_b. Note that the vectors $\vec{\mathbf{y}}_m$ and $\vec{\mathbf{w}}_m$ are not defined anymore, and the noise pdf is written as

$$f_{\mathsf{w}_m(b)}(w_m) = f_{\mathsf{w}(b)}(w_m) = \mathcal{N}(w_m; 0, \sigma_{\mathsf{w}}^2(b)) \,, \quad m \in [M_b] \,. \tag{3.6}$$

Since the noise covariance is diagonal, the estimator derivative in the Onsager term $\frac{1}{M}\mathbf{z}^{(t-1)}(b) \sum_{n=1}^{N} F'\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}^{(t-1)}\right)$ becomes diagonal, i.e.,

$$\frac{1}{M}\mathbf{z}^{(t-1)}(b) \sum_{n=1}^{N} F'\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}^{(t-1)}\right) = \frac{1}{M}\mathbf{z}^{(t-1)}(b) \sum_{n=1}^{N} \frac{d}{du_b} F\left(\vec{\mathbf{u}}_n^{(t-1)}; \Sigma_{\vec{\mathbf{v}}}^{(t-1)}\right) \,, \quad b \in [B] \,.$$

3.4 Priors of Interest

3.4.1 Bernoulli-Gauss Prior

Similar to the scalar case, the BG prior in B dimensions is defined as

$$f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = (1-\epsilon)\delta(\vec{\mathbf{x}}_n) + \epsilon \mathcal{N}(\vec{\mathbf{x}}_n; \mathbf{0}, \Sigma_{\vec{\mathbf{x}}}) \,, \tag{3.7}$$

where ϵ is the nonzero probability, $1-\epsilon$ is the sparsity, and $\Sigma_{\vec{\mathbf{x}}}$ is the covariance matrix of the vector composed of the identically indexed components $\vec{\mathbf{x}}_n = (\mathsf{x}_n(1), \ldots, \mathsf{x}_n(B))^T$. When considering the noisy decoupled measurement with multivariate additive Gaussian noise, i.e.,

$$\vec{\mathbf{u}} = \vec{\mathbf{x}} + \vec{\mathbf{v}} \quad \text{with} \quad \vec{\mathbf{v}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\vec{\mathbf{v}}}) \,,$$

the pdf of $\vec{\mathbf{u}}$ reads

$$f_{\vec{\mathbf{u}}}(\vec{\mathbf{u}}) = (1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}) + \epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}}) \text{ with } \Sigma_{\vec{\mathbf{u}}} = \Sigma_{\vec{\mathbf{x}}} + \Sigma_{\vec{\mathbf{v}}}\,.$$

The MMSE estimator of $\vec{\mathbf{x}}$ given $\vec{\mathbf{u}} = \vec{\mathbf{u}}$ reads

$$\begin{aligned} F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) &= \mathrm{E}\left\{\vec{\mathbf{x}} \mid \vec{\mathbf{u}} = \vec{\mathbf{u}}\right\} \\ &= \frac{\epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}})}{(1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}) + \epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}})}\Sigma_{\vec{\mathbf{x}}}\Sigma_{\vec{\mathbf{u}}}^{-1}\vec{\mathbf{u}} \\ &= \frac{F_N(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}})}{F_D(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}})}\Sigma_{\vec{\mathbf{x}}}\Sigma_{\vec{\mathbf{u}}}^{-1}\vec{\mathbf{u}}\,, \end{aligned} \tag{3.8}$$

with

$$\begin{aligned} F_N(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) &= \epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}})\,, \\ F_D(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) &= (1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}) + \epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}})\,. \end{aligned}$$

Its derivative is the Jacobian matrix

$$\begin{aligned} F'(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) &= \frac{d}{d\vec{\mathbf{u}}^T}F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) \\ &= \frac{1}{F_D}\Bigg(\epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}})\Big(\Sigma_{\vec{\mathbf{x}}}\Sigma_{\vec{\mathbf{u}}}^{-1} - \Sigma_{\vec{\mathbf{x}}}\Sigma_{\vec{\mathbf{u}}}^{-1}\vec{\mathbf{u}}\vec{\mathbf{u}}^T\Sigma_{\vec{\mathbf{u}}}^{-1})\Big) \\ &\quad + F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}})\vec{\mathbf{u}}^T \cdot \Big((1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}; 0, \Sigma_{\vec{\mathbf{v}}})\Sigma_{\vec{\mathbf{v}}}^{-1} + \epsilon\mathcal{N}(\vec{\mathbf{u}}; 0, \Sigma_{\vec{\mathbf{u}}})\Sigma_{\vec{\mathbf{u}}}^{-1}\Big)\Bigg)\,. \end{aligned}$$

3.4.2 Discrete Prior

The discrete prior pdf is a straightforward generalization of the scalar discrete prior and is defined as

$$f_{\mathsf{x}}(\vec{\mathbf{x}}_n) = \sum_{c=1}^{C}\epsilon^{(c)}\delta(\vec{\mathbf{x}}_n - \vec{\mathbf{s}}^{(c)})\,.$$

The symbol alphabet $\mathcal{S} = \{\vec{\mathbf{s}}^{(1)}, \ldots, \vec{\mathbf{s}}^{(C)}\}$ is of size $C = |\mathcal{S}|$ and composed of B-dimensional symbol vectors $\vec{\mathbf{s}}^{(c)} \in \mathbb{R}^B$, and $\epsilon^{(c)}$ ($c \in [C]$) are the individual symbol probabilities, which sum up to 1:

$$P\left\{\vec{\mathbf{x}}_n = \vec{\mathbf{s}}^{(c)}\right\} = \epsilon^{(c)}\,, \quad \text{with} \quad \sum_{c=1}^{C}\epsilon^{(c)} = 1\,.$$

In CS, typically, w.l.o.g. $\mathbf{s}^{(1)} = \mathbf{0}$ and $1 > \epsilon^{(1)} \gg 0$. If there is a dominant symbol vector that is nonzero, the measurement can be transformed into an equivalent measurement following the procedure similar to the mean removal in Section 2.1. Note that this does not influence the performance of BAMP. The discrete prior is a powerful tool for, e.g., telecommunication applications, where the transmit symbols typically constitute a finite alphabet [69, 70]. The MMSE estimator function for the discrete prior pdf given a noisy observation with Gaussian noise with covariance matrix $\Sigma_{\vec{\mathbf{v}}}$ reads

$$F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) = \frac{\sum_{c=1}^{C} \epsilon^{(c)} \vec{\mathbf{s}}^{(c)} \mathcal{N}(\vec{\mathbf{u}}; \vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}})}{\sum_{c=1}^{C} \epsilon^{(c)} \mathcal{N}(\vec{\mathbf{u}}; \vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}})} ,$$

which is essentially a weighted sum of all symbol vectors. Its derivative is

$$\begin{aligned} F'(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) &= \frac{d}{d\vec{\mathbf{u}}^T} F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) \\ &= \frac{1}{F_D(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}})} \sum_{c=1}^{C} \epsilon^{(c)} \mathcal{N}(\vec{\mathbf{u}}; \vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}}) (F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}) - \vec{\mathbf{s}}^{(c)}) (\vec{\mathbf{u}} - \vec{\mathbf{s}}^{(c)})^T \Sigma_{\vec{\mathbf{v}}}^{-1} . \end{aligned}$$

3.5 Soft Information and Reestimation

After meeting the stopping criterion, BAMP and V-BAMP deliver approximate MMSE estimates. It follows that if the prior pdf contains any discrete components, $x_n(b)$ will almost never exactly take on the desired discrete value. In case of the widely employed BG prior, one is often interested in the support $\mathcal{S} = \text{supp}(\mathbf{x}(b))$ $(b \in [B])$ of the signal, i.e., the set of (non)zero components/indices. And while approximately $(1 - \epsilon)N$ components are expected to be 0, the (V)BAMP estimate $\hat{\mathbf{x}}(b)$ $(b \in [B])$ contains in general no zeros. Furthermore, in case of the discrete prior, one needs to transform the obtained continuous values to discrete values in order to obtain a final valid estimate. Thus, post-processing is necessary.

3.5.1 Expectation-Maximization-based Classification

It is clear that if $\|\vec{\mathbf{u}}_n^{(t)}\|_2$ is relatively large, based on the decoupled measurement model one can be confident that $\vec{\mathbf{x}}_n$ is a nonzero vector and hence $n \in \mathcal{S}$. On the other hand, if $\|\vec{\mathbf{u}}_n^{(t)}\|_2$ is relatively small, one cannot be sure whether $\hat{\vec{\mathbf{x}}}_n^{(t)}$ is a noisy estimate of $\vec{\mathbf{x}}_n = 0$ or a (noisy) estimate of a small but nonzero $\vec{\mathbf{x}}_n$. A theoretically sound way of (soft) clustering vectors (numbers) that are assumed to come from different distributions is the EM algorithm [67, 79]. For Gaussian mixture distributions, the EM algorithm not only classifies the vectors (E-step), but also finds its parameters

(mean, (co-)variance, occurrence probability) in the M-step. Since V-BAMP already delivers those parameters, only a single E-step is necessary for classification. The E-step calculates the so called *responsibilities*. The responsibility is a measure of how well an observation is explained by a certain component distribution. By deciding for the component distribution with the highest responsibility one achieves statistically optimal classification [67] and quantization to the desired discrete values. Furthermore, by saving the soft information (i.e., the responsibilities) further post-processing is possible. This has been exploited in a number of works [47, 48, 80].

Bernoulli-Gauss prior

After including the knowledge that the event $n \notin \mathcal{S}$ has probability $(1-\epsilon)$, while $n \in \mathcal{S}$ has probability ϵ, one can write the distribution of $\vec{\mathbf{u}}_n^{(t)}$ as

$$f_{\vec{\mathbf{u}}_n^{(t)}}(\vec{\mathbf{u}}_n^{(t)}) = (1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)}) + \epsilon\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)} + \Sigma_{\vec{\mathbf{x}}})\,.$$

Formally, $\vec{\mathbf{u}}_n^{(t)}$ comes from one of the two distributions:

$$\vec{\mathbf{u}}_n^{(t)} \sim \begin{cases} \mathcal{N}(\mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)}) & \text{if } n \notin \mathcal{S}\,, \\ \mathcal{N}(\mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)} + \Sigma_{\vec{\mathbf{x}}}) & \text{if } n \in \mathcal{S}\,, \end{cases}$$

i.e., $\vec{\mathbf{u}}_n^{(t)}$ contains only *effective noise*, or *signal* plus *effective noise*. The E-step calculates responsibilities as

$$\begin{aligned} \rho(n \notin \mathcal{S}) &= \mathrm{P}\{n \notin \mathcal{S} \mid \vec{\mathbf{u}}_n^{(t)} = \vec{\mathbf{u}}_n^{(t)}\} = \frac{1}{Z}(1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)})\,, \\ \rho(n \in \mathcal{S}) &= \mathrm{P}\{n \in \mathcal{S} \mid \vec{\mathbf{u}}_n^{(t)} = \vec{\mathbf{u}}_n^{(t)}\} = \frac{1}{Z}\epsilon\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)} + \Sigma_{\vec{\mathbf{x}}})\,, \end{aligned}$$

with Z being a normalization constant.

Discrete prior

When the prior contains only discrete values (symbols), $\vec{\mathbf{x}}_n^{(t)}$ is expected to be close to $\vec{\mathbf{u}}_n^{(t)}$. Inserting the prior symbol probabilities $\epsilon^{(c)}$, the distribution of $\vec{\mathbf{u}}_n^{(t)}$ reads

$$f_{\vec{\mathbf{u}}_n^{(t)}}(\vec{\mathbf{u}}_n^{(t)}) = \sum_{c=1}^{C} \epsilon^{(c)}\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}}^{(t)})\,.$$

Formally,

$$\vec{\mathbf{u}}_n^{(t)} \sim \mathcal{N}(\vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}}^{(t)}) \text{ if } \vec{\mathbf{x}}_n = \vec{\mathbf{s}}^{(c)} .$$

The E-step calculates responsibilities as

$$\rho(\vec{\mathbf{x}}_n = \vec{\mathbf{s}}^{(c)}) = \mathrm{P}\{\vec{\mathbf{x}}_n = \vec{\mathbf{s}}^{(c)} \mid \vec{\mathbf{u}}_n^{(t)} = \vec{\mathbf{u}}_n^{(t)}\} = \frac{1}{Z}\epsilon^{(c)}\mathcal{N}(\vec{\mathbf{u}}_n^{(t)}; \vec{\mathbf{s}}^{(c)}, \Sigma_{\vec{\mathbf{v}}}^{(t)})$$

$c \in [C]$, with Z being a normalization constant. Note that if the symbol probabilities $\epsilon^{(c)}$ of the discrete prior are identical, i.e., $\epsilon^{(c)} = \epsilon$, $c \in [C]$, then the C responsibilities are inversely (exponentially) proportional to the distance of the decoupled measurement to each of the C symbol vectors, i.e., $\vec{\mathbf{u}}_n^{(t)} - \vec{\mathbf{s}}^{(c)}$. That is, the classification based on the E-step responsibilities is identical to nearest neighbor quantization.

3.5.2 Reestimation

Suppose the signal prior pdf contains discrete components as well as continuous components. With the E-step described above, one is able to fine tune the estimates $\hat{\vec{\mathbf{x}}}_n$ on a subset of the indices $[N]$ and assign them to discrete values from the prior pdf. The values classified as *coming from* the continuous component distribution(s) are unaltered. Let us collect the indices of these values into the set $\mathcal{S}_c$ (c for $c$ontinuous) and the remaining indices into the set $\mathcal{S}_d = [N] \setminus \mathcal{S}_c$ (d for $d$iscrete). Then the measurements can be written as

$$\mathbf{y}(b) = \mathbf{A}(b)_{\mathcal{S}_c}\mathbf{x}(b)_{\mathcal{S}_c} + \mathbf{A}(b)_{\mathcal{S}_d}\mathbf{x}(b)_{\mathcal{S}_d} + \mathbf{r}(b) ,$$

where $\mathbf{r}(b)$ denotes the residual noise. Rearranging the terms in the above equation results in

$$\underbrace{\mathbf{y}(b) - \mathbf{A}(b)_{\mathcal{S}_d}\hat{\mathbf{x}}(b)_{\mathcal{S}_d}}_{\bar{\mathbf{y}}(b)} = \underbrace{\mathbf{A}(b)_{\mathcal{S}_c}}_{\bar{\mathbf{A}}(b)}\underbrace{\hat{\mathbf{x}}(b)_{\mathcal{S}_c}}_{\bar{\mathbf{x}}(b)} + \mathbf{r}(b) ,$$

where the left side is known due to the assumption that the discrete values have been assigned correctly by the E-step. Now one can write the reduced measurement

$$\bar{\mathbf{y}}(b) = \bar{\mathbf{A}}(b)\bar{\mathbf{x}}(b) + \mathbf{r}(b) , \tag{3.9}$$

where the dimensions of the reduced measurement matrix $\bar{\mathbf{A}}(b)$ are (depending on the fraction of excluded discrete components) much more favorable than those of $\mathbf{A}(b)$. In some cases, $\bar{\mathbf{A}}(b)$ might even be overdetermined. The reduced measurement (3.9) allows for *reestimation* of a subset of the components, for which various methods are

available: V-BAMP, least squares estimation etc. Note that the EM algorithm and reestimation can be combined in several ways, even in an iterative manner, in order to exploit soft information and achieve more accurate estimation depending on the demands [72].

3.6 State Evolution

The SE presented in Section 2.3.1 for BAMP has been extended to the MMV and DCS scenarios in, e.g., [81]. It allows us to analytically describe the behavior of V-BAMP. (We point the interested reader to the fact that in [81] the Onsager term is defined incorrectly; nonetheless, the presentation of the multivariate SE is correct.) The SE equation allows us to track the evolution of the effective noise covariance (its state) across iterations in an iterative manner:

$$\begin{aligned}\Sigma_{\vec{\mathbf{v}}}{}^{(t+1)} &= S(\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) \\ &= \begin{cases} \Sigma_{\vec{\mathbf{w}}} + \frac{1}{R}\,\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\} & \text{for MMV}\,, \\ \operatorname{diag}\left(\Sigma_{\vec{\mathbf{w}}} + \frac{1}{R}\,\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\}\right) & \text{for DCS}\,, \end{cases}\end{aligned} \tag{3.10}$$

for a general signal prior $\vec{\mathbf{x}} \sim f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n)$ and estimator $F(\vec{\mathbf{u}}_n^{(t)};\Sigma_{\vec{\mathbf{v}}})$, with $\vec{\mathbf{v}} \sim \mathcal{N}(\mathbf{0},\Sigma_{\vec{\mathbf{v}}}^{(t)})$. (Note that the SE equations are valid even if the estimator $F()$ is not the MMSE estimator of $\vec{\mathbf{x}}$.) The state of V-BAMP is B-dimensional for DCS and in general $B(B+1)/2$-dimensional for MMV (since the covariance matrix is symmetric). From (3.10) the MSE prediction follows as

$$\begin{aligned}\operatorname{Cov}(\vec{\mathbf{u}}_n^{(t)} - \vec{\mathbf{x}}_n) &= \Sigma_{\vec{\mathbf{v}}}^{(t)}\,, \\ \widehat{\mathrm{MSE}}(\hat{\mathbf{x}}^{(t)}(b),\mathbf{x}(b)) &= R(\Sigma_{\vec{\mathbf{v}}}^{(t)} - \Sigma_{\vec{\mathbf{w}}})_{b,b}\,.\end{aligned}$$

Details on the numerical evaluation of (3.10) can be found in Appendix D.

Discussion

In Figure 3.1 and Figure 3.2, the experiments from Section 2.3.1 were repeated (corresponding to Figure 2.5 and Figure 2.6, respectively) with the $B = 5$-dimensional BG prior

$$f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = 0.9\,\delta(\vec{\mathbf{x}}_n) + 0.1\,\mathcal{N}(\vec{\mathbf{x}}_n;\mathbf{0},\mathbf{I}_5)$$

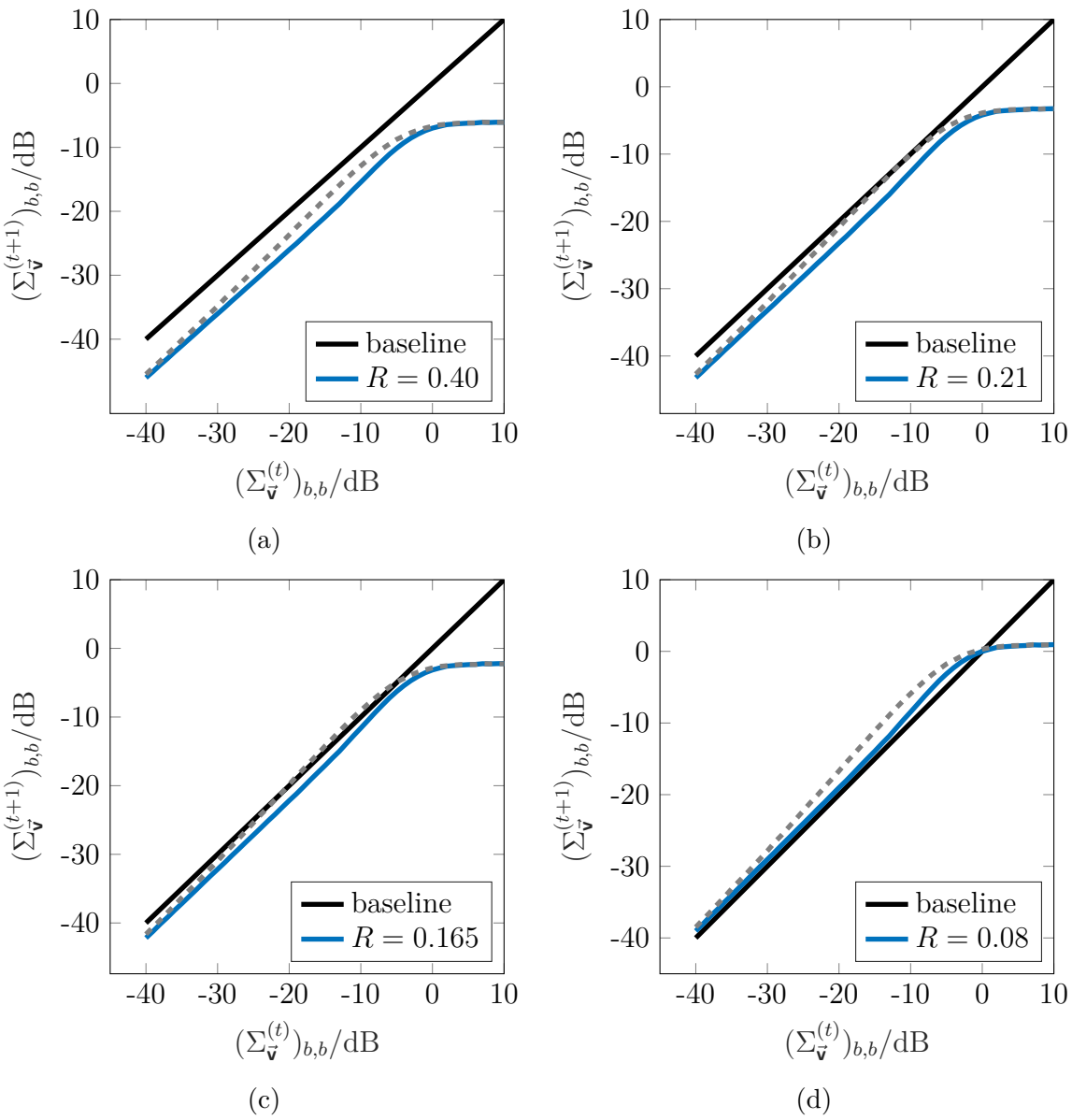


Fig. 3.1 Empirical noiseless SE curves for the 5-D multivariate BG prior and MMSE estimator obtained from Monte Carlo simulation (dimension $N = 1000$, $\epsilon = 0.1$, $\Sigma_{\vec{\mathbf{w}}} = \mathbf{0}$). The SE curves for the scalar case ($B = 1$, identical parameters) are shown in gray dashed lines for comparison.

and the Bayesian setting, i.e., the corresponding MMSE estimator. In this case the MMV and DCS SE are identical (the proof is elaborated in Appendix C). Moreover, when both the signal covariance and the additive noise covariance are diagonal and constant along their diagonals, the effective noise covariance is diagonal as well and defined by a single parameter, $(\Sigma_{\vec{\mathbf{v}}}^{(t)})_{b,b}$ (identical for $b \in [B]$). Thus, it is possible to

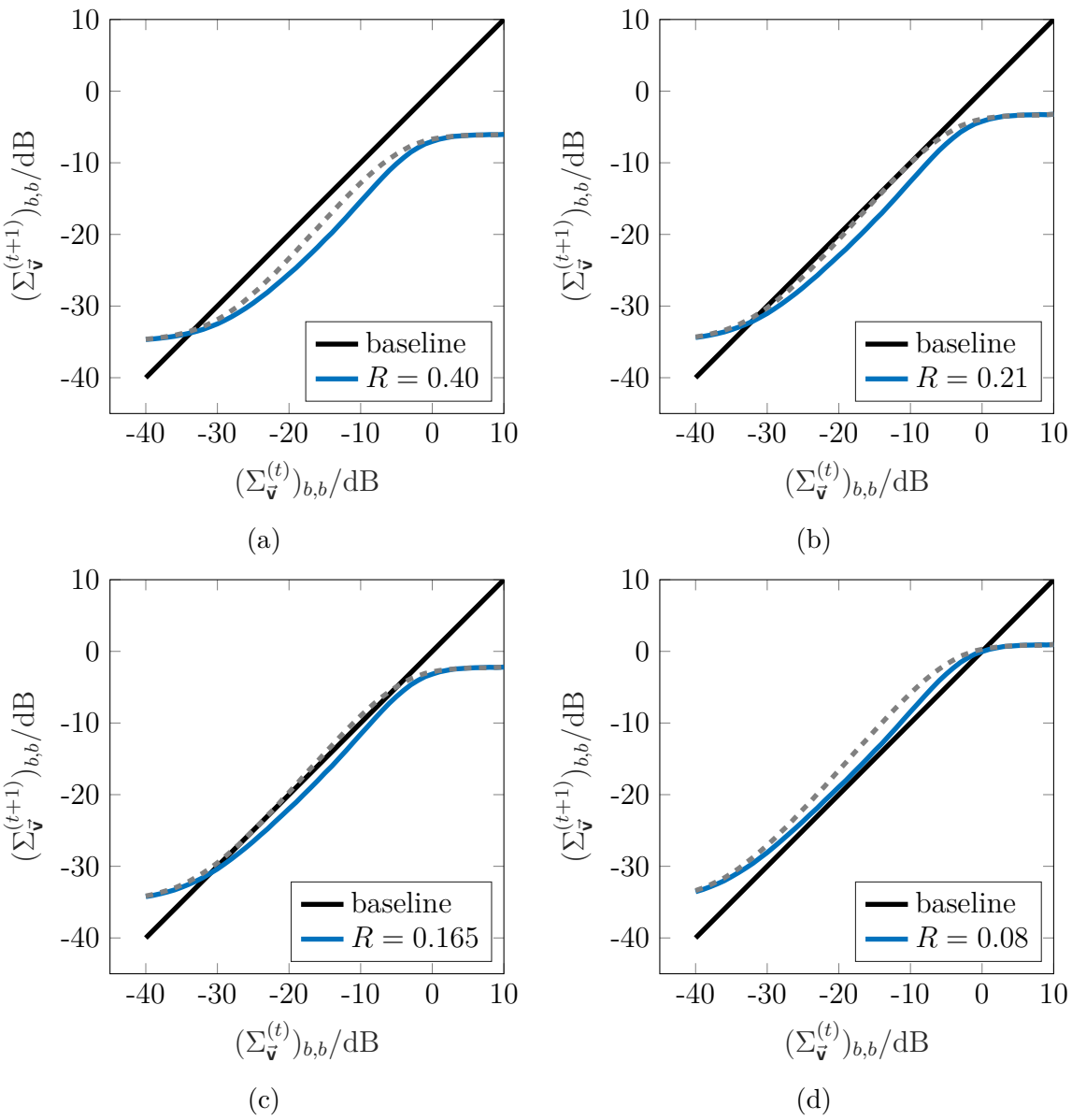


Fig. 3.2 Empirical noisy SE curves for the 5-D multivariate BG prior and MMSE estimator obtained from Monte Carlo simulation (dimension $N = 1000$, $\epsilon = 0.1$, $\Sigma_{\vec{\mathbf{w}}} = -35\text{dB}\,\mathbf{I}$). The SE curves for the scalar case ($B = 1$, identical parameters) are shown in gray dashed lines for comparison.

visualize the SE with a 1-dimensional curve instead of a $B(B+1)$-dimensional function. Since V-BAMP exploits the knowledge of the common support of the measured vectors, we expect a drop of the PT rate in the noiseless case relative to the $B = 1$ case. Observe that the SE curves drop, i.e., in the noiseless case V-BAMP converges faster and its PT rate lower, while in the noisy case the reached effective noise variances

are lower. Note that the nonzero signal components are uncorrelated, i.e., the only information that is common among the 5 vectors is their support. In Figure 3.3, the gain from increasing the number of jointly sparse vectors is investigated by plotting the empirical noiseless SE curve for $\epsilon = 0.1$, $B = 1$, $B = 2$, and $B = 10$, with parameters to $\Sigma_{\vec{\mathbf{w}}} = \mathbf{0}$, rate $R = 0.21$ (which is approximately R_{PT} for $B = 1$), and identical signal powers, i.e., $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}$. Observe that while the gain from additional jointly sparse vectors seems minor, there is a significant difference for larger B values (discussed in the following). In Figure 3.3b, the correlation coefficient between each pair of the Gaussian components was set to 0.9, i.e., $(\Sigma_{\vec{\mathbf{x}}})_{b,b'} = C = 0.9$ for $b \neq b'$. (Note that a correlation coefficient 1 in the noiseless case corresponds to the single measurement vector problem for MMV and doubling the measurement rate R for DCS. Further discussion on the effect of signal correlation can be found in Section 3.8.) We highlight that in this case the SE is not described by a single quantity (the diagonal $(\Sigma_{\vec{\mathbf{v}}}^{(t)})_{b,b}$) anymore because the off-diagonal elements of the effective noise covariance ($(\Sigma_{\vec{\mathbf{v}}}^{(t)})_{b,b'}$, $b \neq b'$) are in general nonzero. Nonetheless, plotting the evolution of the diagonal elements gives insight into the MSE performance of V-BAMP. Observe that the gain from additional jointly sparse vectors is stronger relative to the uncorrelated signal case, i.e., there is a considerable sampling rate and noise regime in which V-BAMP will be successful only by exploiting information from additional vectors. In Figure 3.4, the effect of the correlation coefficient C on V-BAMP is investigated. The reduced SE curves representing the evolution of the diagonal elements $(\Sigma_{\vec{\mathbf{v}}}^{(t)})_{b,b}$ for $\epsilon = 0.1$, $B = 10$ jointly sparse vectors, $R = 0.21$, and no additive noise are plotted for different correlation coefficients C, i.e., $(\Sigma_{\vec{\mathbf{x}}})_{b,b'} = C$ $(b' \neq b)$ while $(\Sigma_{\vec{\mathbf{x}}})_{b,b} = 1$. Observe that increasing correlation results in better performance (in an MSE or convergence speed sense or both) as the reduced SE curve drops.

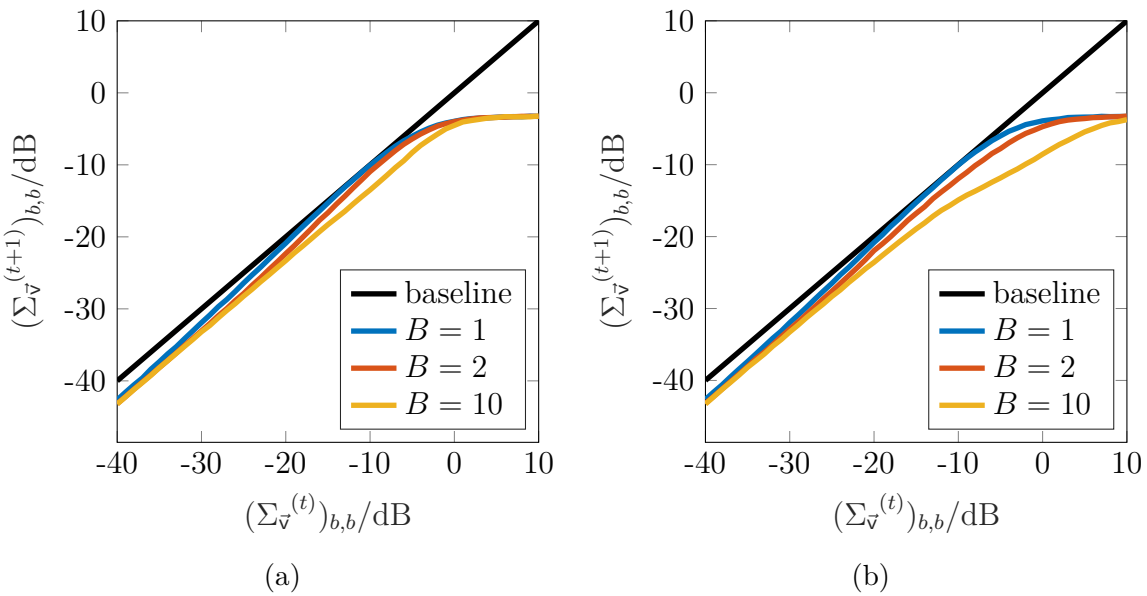


Fig. 3.3 Empirical SE curves for different number of (a) uncorrelated (b) highly correlated identically BG distributed components (jointly sparse vectors). In the correlated case only the diagonal effective noise variance evolution is plotted.

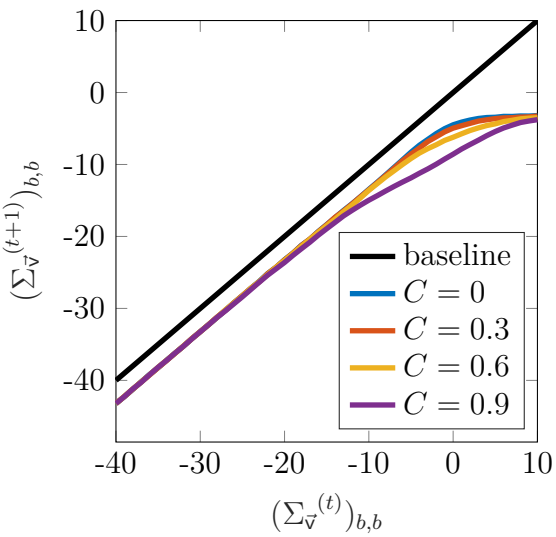


Fig. 3.4 Empirical SE curves for different correlation coefficients C for $B = 10$ identically BG distributed components (jointly sparse vectors, $\epsilon = 0.1$, $N = 1000$, $(\Sigma_{\vec{\mathbf{x}}})_{b,b'} = C$ for $b' \neq b$, $(\Sigma_{\vec{\mathbf{x}}})_{b,b} = 1$). Only the diagonal effective noise variance evolution is plotted.

Noiseless PT curves obtained from the SE MSE prediction are plotted in Figure 3.5, for $B = 1, 2, 3$-dimensional isotropic uncorrelated BG prior. Even though the advantage of additional (uncorrelated) jointly sparse vectors seems minor in the SE (Figure 3.3a), the differences are pronounced at rates at the PT rate (and also with

increasing additive noise and signal correlation). Also note that the PT happens only in the CS regime, i.e., where $\epsilon \ll 1$. As B increases, the sudden transition requires smaller ϵ: while at $\epsilon = 0.2$ for $B = 1$ and $B = 2$ (Figures 3.5a and 3.5b) there is still a sudden drop in the MSE, for $B = 3$ at $\epsilon = 0.2$ the transition is already smooth. Further investigation on the PT of jointly sparse CS can be found in Section 3.9.

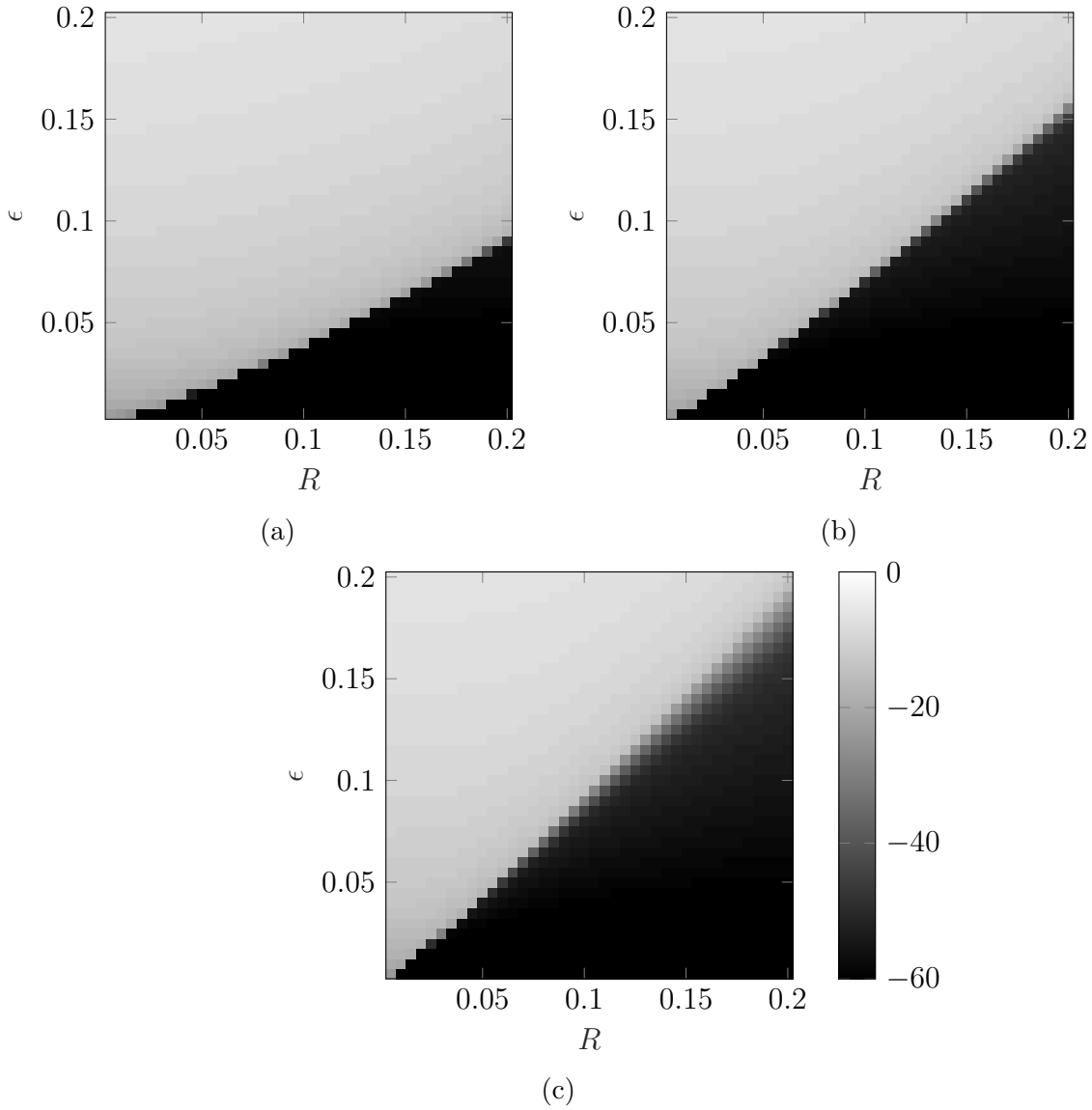


Fig. 3.5 MSE of V-BAMP as predicted by SE, as a function of the sampling rate R and the nonzero probability ϵ with (a) $B = 1$-D, (b) $B = 2$-D, (c) $B = 3$-D BG signal prior with $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}$. The shade corresponds to $\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}(b), \mathbf{x}(b))/\mathrm{dB}$ as $t \to \infty$, identically for $b \in [B]$. The PT curve is what separates the dark region (corresponding to successful recovery) from the bright region (unsuccessful recovery) in the low ϵ regime.

Note that above the SE and PT analysis are precisely valid only for the signals with (multivariate) BG prior and its corresponding MMSE estimator. The SE curves for other signals and other estimators might look differently. However, in the CS regime, where the sparsity is high, the deviation in terms of the fixed points of the SE equation and of the PT curves is expected to be minor.

3.7 Joint Diagonalization for MMV

3.7.1 Joint Diagonalization of the Measurements

The presented V-BAMP algorithm for DCS and MMV is a powerful tool that can deal with arbitrary signal and noise correlations $\Sigma_{\vec{\mathsf{x}}}$ and $\Sigma_{\vec{\mathsf{w}}}$. In the DCS scenario the effective noise covariance $\Sigma_{\vec{\mathsf{v}}}^{(t)}$ is always a diagonal matrix due to the independence of the measurement matrices $\mathbf{A}(b)$, $b \in [B]$. In contrast, in the MMV case in general $\Sigma_{\vec{\mathsf{v}}}^{(t)}$ will be nondiagonal, even if the signal prior and the additive noise are uncorrelated themselves, i.e., if $\Sigma_{\vec{\mathsf{x}}}$ and $\Sigma_{\vec{\mathsf{w}}}$ are diagonal. This means $\mathcal{O}(B^2)$ SNR relations in the decoupled measurements $\vec{\mathbf{u}} = \vec{\mathbf{x}} + \vec{\mathbf{v}}$: each $\mathsf{x}_n(b)$ is correlated with all $\mathsf{x}_n(b')$ and is thus influenced simultaneously by all effective noise components $\mathsf{v}_m(b')$ ($b' \in [B], n \in [N], m \in [M]$). In other words, the V-BAMP system is described by $B(B+1)/2$ states, even if only the B effective noise variances on the diagonal are of interest after convergence (since these store the MSE estimates). Next we show that any MMV measurement of the form (3.1) or (3.4) can be transformed into one with diagonal signal and diagonal noise covariance.

Our aim is to transform (3.1) by a nonsingular matrix $\mathbf{T}$ such that $\mathrm{Cov}\{\mathbf{T}\vec{\mathbf{x}}_n\} = \mathbf{T}\,\mathrm{Cov}\{\vec{\mathbf{x}}_n\}\mathbf{T}^T = \mathbf{T}\epsilon\Sigma_{\vec{\mathsf{x}}}\mathbf{T}^T$ and $\mathrm{Cov}\{\mathbf{T}\vec{\mathbf{w}}_m\} = \mathbf{T}\Sigma_{\vec{\mathsf{w}}}\mathbf{T}^T$ are both diagonal:

$$\begin{aligned}\Big(\mathbf{y}(1),\dots,\mathbf{y}(B)\Big)\mathbf{T}^T &= \Big(\mathbf{A}(1)\mathbf{x}(1),\dots,\mathbf{A}(B)\mathbf{x}(B)\Big)\mathbf{T}^T + \Big(\mathbf{w}(1),\dots,\mathbf{w}(B)\Big)\mathbf{T}^T \qquad (3.11)\\ &= \mathbf{A}\Big(\mathbf{x}(1),\dots,\mathbf{x}(B)\Big)\mathbf{T}^T + \Big(\mathbf{w}(1),\dots,\mathbf{w}(B)\Big)\mathbf{T}^T\\ &= \mathbf{A}\mathbf{X}\mathbf{T}^T + \mathbf{W}\mathbf{T}^T\\ &\Updownarrow\\ \tilde{\vec{\mathbf{y}}}_m = \mathbf{T}\vec{\mathbf{y}}_m\,,\quad \tilde{\vec{\mathbf{x}}}_n &= \mathbf{T}\vec{\mathbf{x}}_n\,,\quad \tilde{\vec{\mathbf{w}}}_m = \mathbf{T}\vec{\mathbf{w}}_m\,. \qquad (3.12)\end{aligned}$$

Remember that $\Sigma_{\vec{\mathsf{x}}}$ denotes the covariance of $\vec{\mathbf{x}}$ given that it is nonzero, i.e., $\Sigma_{\vec{\mathsf{x}}} = \mathrm{Cov}\{\vec{\mathbf{x}}_n \mid \vec{\mathbf{x}}_n \neq \mathbf{0}\}$. Under the assumption that the covariance matrices $\Sigma_{\vec{\mathsf{x}}}, \Sigma_{\vec{\mathsf{w}}}$ are full rank and using the fact that covariance matrices are symmetric and positive definite

Algorithm 3 Joint diagonalization transformation

1: Given $\Sigma_{\vec{\mathbf{x}}}, \Sigma_{\vec{\mathbf{w}}}$
2: find $\mathbf{P}$ such that $\mathbf{P}\mathbf{P}^T = \Sigma_{\vec{\mathbf{w}}}$
3: $\mathbf{G} = \mathbf{P}^{-1}\Sigma_{\vec{\mathbf{x}}}\mathbf{P}^{-T}$
4: find eigendecomposition $\mathbf{Q}_G\mathbf{\Lambda}_G\mathbf{Q}_G^{-1} = \mathbf{G}$
5: $\mathbf{T} = \mathbf{\Lambda}_G^{-1/2}\mathbf{Q}_G^T\mathbf{P}^{-1}$

and [82, Thm. 7.6.1.], $\mathbf{T}$ exists and can be computed using Algorithm 3. Here, $\mathbf{Q}_G$ is the matrix of orthonormal eigenvectors, and $\mathbf{\Lambda}_G$ is a diagonal matrix containing the eigenvalues. Note that if the goal is only joint diagonalization, infinitely many matrices $\mathbf{T}$ are suitable. In particular, for any diagonal matrix $\mathbf{D}$, $\mathbf{T} = \mathbf{D}\mathbf{Q}_G^T\mathbf{P}^{-1}$ is suitable for joint diagonalization. However, the choice $\mathbf{D} = \mathbf{\Lambda}_G^{-1/2}$ (Step 5 in Algorithm 3) results in

$$\begin{aligned}\mathrm{Cov}\{\vec{\mathbf{x}}_n\mathbf{T}\} &= \epsilon\Sigma_{\tilde{\vec{\mathbf{x}}}} = \epsilon\mathbf{I}_B\,,\\ \mathrm{Cov}\{\vec{\mathbf{w}}_m\mathbf{T}\} &= \Sigma_{\tilde{\vec{\mathbf{w}}}} = \mathrm{diag}\left(\frac{1}{\mathrm{SNR}(1)},\ldots,\frac{1}{\mathrm{SNR}(B)}\right)\,,\end{aligned}$$

where the now independent inverse SNRs of the B measurements are carried directly in the transformed noise covariance matrix $\Sigma_{\tilde{\vec{\mathbf{w}}}}$ (and the off-diagonals are zero). In the uncorrelated signal and uncorrelated noise case the SNR per channel is defined as

$$\mathrm{SNR}(b) = \mathrm{E}_{\mathbf{x},\mathbf{w}}\left\{\frac{\|\mathbf{A}(b)\tilde{\mathbf{x}}(b)\|_2^2}{\|\tilde{\mathbf{w}}(b)\|_2^2}\right\} = \frac{\epsilon(\Sigma_{\tilde{\vec{\mathbf{x}}}})_{b,b}}{(\Sigma_{\tilde{\vec{\mathbf{w}}}})_{b,b}}\,.$$

If one wishes to apply V-BAMP to the transformed measurement model (3.11), the change in the prior pdfs has to be taken into account, which might be difficult for general prior pdfs. That is, the MMSE estimator (3.5) and its derivative will have a new form.

Joint Diagonalization Algorithm

The matrix $\mathbf{T}$ given by Algorithm 3 performs noise whitening and signal decorrelation simultaneously. While simply $\mathbf{T}\Sigma_{\vec{\mathbf{x}}}\mathbf{T}^T = \mathbf{I}_B$, $\mathbf{T}\Sigma_{\vec{\mathbf{w}}}\mathbf{T}^T$ depends on $\Sigma_{\vec{\mathbf{w}}}$ and $\Sigma_{\vec{\mathbf{w}}}$ in a nontrivial way. However, there are two important special cases that give insight into how the resulting $\Sigma_{\tilde{\vec{\mathbf{w}}}}$ arises.

- $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_B$ and $\Sigma_{\vec{\mathbf{w}}}$ is arbitrary (full rank). Through simple calculation $\Sigma_{\tilde{\vec{\mathbf{w}}}} = \Lambda_{\vec{\mathbf{w}}} = \mathrm{diag}(\lambda_{\vec{\mathbf{w}}}(1),\ldots,\lambda_{\vec{\mathbf{w}}}(B))$, i.e., the diagonal matrix with the B eigenvalues of $\Sigma_{\vec{\mathbf{w}}}$.

That is, the underlying inverse SNRs correspond exactly to the eigenvalues of the noise covariance matrix, as $\Sigma_{\tilde{\vec{\mathsf{x}}}} = \Sigma_{\vec{\mathsf{x}}} = \mathbf{I}_B$.

- $\Sigma_{\vec{\mathsf{w}}} = \mathbf{I}_B$ and $\Sigma_{\vec{\mathsf{x}}}$ is arbitrary (full rank). It follows that $\Sigma_{\tilde{\vec{\mathsf{x}}}} = \mathbf{I}_B$ and $\Sigma_{\tilde{\vec{\mathsf{w}}}} = \Lambda_{\vec{\mathsf{x}}}^{-1} = \mathrm{diag}(\frac{1}{\lambda_{\vec{\mathsf{x}}}(1)}, \ldots, \frac{1}{\lambda_{\vec{\mathsf{x}}}(B)})$, i.e., the diagonal matrix composed of the B reciprocal eigenvalues of the signal covariance matrix. That is, the eigenvalues of the nonzero signal covariance matrix correspond exactly to the SNR values.

In any other case, the SNRs depend in a nontrivial way on the covariance matrices. They can be, however, bounded by $\frac{\max(\lambda_{\vec{\mathsf{x}}})}{\min(\lambda_{\vec{\mathsf{w}}})} \geq \mathrm{SNR}(b) \geq \frac{\min(\lambda_{\vec{\mathsf{x}}})}{\max(\lambda_{\vec{\mathsf{w}}})}$, $b \in [B]$.

Parameter Estimation

When the measurement matrices $\mathbf{A}(b)$ ($b \in [B]$) have normalized columns, by simple calculation

$$\mathrm{Cov}\left(\vec{\mathbf{y}}_m\right) = \begin{cases} \Sigma_{\vec{\mathsf{w}}} + \frac{1}{R}\,\mathrm{Cov}\{\vec{\mathbf{x}}_n\} & \mathrm{MMV}\,, \\ \Sigma_{\vec{\mathsf{w}}} + \frac{1}{R}\,\mathrm{diag}\left(\{\mathrm{Var}\{\mathsf{x}_n(b)\}_{b=1,\ldots,B}\}\right) & \mathrm{DCS}\,. \end{cases}$$

That is, either the noise or the signal covariance has to be known in order to estimate the other via the observed (sample) covariance $\mathrm{Cov}\left(\vec{\mathbf{y}}_m\right)$. We refer the interested reader to the EM AMP approach introduced in [72], which could be applied to estimate the unknown parameters during iterations, as long as the measured signal comes from a (Bernoulli-)Gaussian mixture.

3.7.2 Equivalence of the Transformed Model

Next we discuss the fact that both MMV V-BAMP and its SE are equivariant w.r.t. (invertible) linear transformations of the input as long as $F()$ is the MMSE estimator matched for the signal prior $f_{\vec{\mathsf{x}}}(\vec{\mathbf{x}}_n)$. That is, transforming the system once, then iterating V-BAMP on the transformed measurement, and ultimately transforming back the variables of interest is equivalent to iterating V-BAMP on the original measurement.

Theorem 2 *Algorithm 2 for MMV and its SE are equivariant w.r.t. invertible linear transformations if $F()$ is the MMSE estimator matched for the signal prior $f_{\vec{\mathsf{x}}}(\vec{\mathbf{x}}_n)$. That is, for any nonsingular $\mathbf{T}$:*

1. *If one iteration of Algorithm 2 maps $(\vec{\mathbf{y}}_m, \hat{\vec{\mathbf{x}}}_n^{(t)}, \vec{\mathbf{r}}_m^{(t)}, \Sigma_{\vec{\mathbf{v}}}^{(t)}) \to (\hat{\vec{\mathbf{x}}}_n^{(t+1)}, \vec{\mathbf{r}}_m^{(t+1)}, \Sigma_{\vec{\mathbf{v}}}^{(t+1)})$, then it maps $(\mathbf{T}\vec{\mathbf{y}}_m, \mathbf{T}\hat{\vec{\mathbf{x}}}_n^{(t)}, \mathbf{T}\vec{\mathbf{r}}_m^{(t)}, \mathbf{T}\Sigma_{\vec{\mathbf{v}}}^{(t)}\mathbf{T}^T) \to (\mathbf{T}\hat{\vec{\mathbf{x}}}_n^{(t+1)}, \mathbf{T}\vec{\mathbf{r}}_m^{(t+1)}, \mathbf{T}\Sigma_{\vec{\mathbf{v}}}^{(t+1)}\mathbf{T}^T)$, $\forall m, n$.*

2. The following transformed SE equation holds

$$\mathbf{T}\Sigma_{\vec{\mathbf{v}}}^{(t+1)}\mathbf{T}^T = \mathbf{T}\Sigma_{\vec{\mathbf{w}}}\mathbf{T}^T + \frac{1}{R}\,\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\mathbf{T}(\vec{\mathbf{x}}+\vec{\mathbf{v}});\mathbf{T}\Sigma_{\vec{\mathbf{v}}}{}^{(t)}\mathbf{T}^T) - \mathbf{T}\vec{\mathbf{x}}\rangle\right\} \tag{3.13}$$

$$\Updownarrow$$

$$\Sigma_{\vec{\mathbf{v}}}^{(t+1)} = \Sigma_{\vec{\mathbf{w}}} + \mathbf{T}^{-1}\frac{1}{R}\,\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\mathbf{T}(\vec{\mathbf{x}}+\vec{\mathbf{v}});\mathbf{T}\Sigma_{\vec{\mathbf{v}}}{}^{(t)}\mathbf{T}^T) - \mathbf{T}\vec{\mathbf{x}}\rangle\right\}\mathbf{T}^{-T}\,. \tag{3.14}$$

The proofs of Theorem 2 is elaborated in Appendix B. An important consequence is that if Algorithm 2 converges to $\hat{\mathbf{x}}(b)$ with inputs $\mathbf{y}(b), \Sigma_{\vec{\mathbf{x}}}, \Sigma_{\vec{\mathbf{w}}}$, and it converges to $\tilde{\hat{\mathbf{x}}}(b)$ with inputs $\tilde{\mathbf{y}}(b), \Sigma_{\tilde{\vec{\mathbf{x}}}}, \Sigma_{\tilde{\vec{\mathbf{w}}}}$, then $\mathbf{T}^{-1}\tilde{\hat{\vec{\mathbf{x}}}}_n = \hat{\vec{\mathbf{x}}}_n$ ($b \in [B]$, $n \in [N]$).

3.7.3 Bernoulli-Gauss Prior

As discussed in Section 3.7.1, even if the signal prior and the noise prior possess diagonal covariance structure, i.e., $\Sigma_{\vec{\mathbf{x}}}$ and $\Sigma_{\vec{\mathbf{w}}}$ are diagonal matrices (and so is $\Sigma_{\vec{\mathbf{v}}}^{(0)}$), as V-BAMP iterates, the effective noise covariance $\Sigma_{\vec{\mathbf{v}}}^{(t)}$ ($t > 0$) does not preserve this property in general. The intuition behind this is that while the variable $\Sigma_{\vec{\mathbf{v}}}^{(t)}$ is solely a second moment, the MMSE estimator $F(\vec{\mathbf{u}}_n^{(t)}, \Sigma_{\vec{\mathbf{v}}}^{(t)})$ takes the full pdf of the signal into account. Remember that the Gaussian probability distribution is fully characterized by its first two moments, the mean and the covariance. Moreover, the generalized pdf described by a centered Dirac ($\delta(\vec{\mathbf{x}}_n)$) has zero covariance. Taking a linear combination of a Dirac function and a Gaussian pdf is thus also fully described by its first two moments (and the weights of the linear combination, i.e., the nonzero probability ϵ). To summarize, one can expect that for the BG prior V-BAMP and its SE preserve the (once given) diagonal property of the effective noise covariance $\Sigma_{\vec{\mathbf{v}}}^{(t)}$. For the BG prior, after applying the transformation $\mathbf{T}$, the equivalent measurement model becomes

$$\tilde{\mathbf{y}}(b) = \mathbf{A}(b)\tilde{\mathbf{x}}(b) + \tilde{\mathbf{w}}(b)\,. \tag{3.15}$$

In the MMV scenario, $\mathbf{T}$ can be compactly incorporated as

$$\mathbf{Y}\mathbf{T}^T = \mathbf{A}\mathbf{X}\mathbf{T}^T + \mathbf{W}\mathbf{T}^T\,.$$

The transformed pdfs read

$$f_{\tilde{\vec{\mathbf{x}}}}(\tilde{\vec{\mathbf{x}}}_n) = (1-\epsilon)\delta(\tilde{\vec{\mathbf{x}}}_n) + \epsilon\mathcal{N}(\tilde{\vec{\mathbf{x}}}_n; \mathbf{0}, \mathbf{I}_B)\,, \tag{3.16}$$

$$f_{\tilde{\vec{\mathbf{w}}}}(\tilde{\vec{\mathbf{w}}}_m) = \mathcal{N}(\tilde{\vec{\mathbf{w}}}_m; \mathbf{0}, \Sigma_{\tilde{\vec{\mathbf{w}}}})\,, \tag{3.17}$$

i.i.d. over m, n, and with diagonal $\Sigma_{\tilde{\vec{\mathbf{w}}}}$. That is, a BG prior in the transformed domain is retained but with uncorrelated components. This is peculiar to the BG prior: in general the joint distribution will have a different form to the original one.
Consider the SE equation (3.10) that describes the expected evolution of the effective noise covariance across the V-BAMP iterations. In the MMV scenario, apart from some special cases, even if $\Sigma_{\vec{\mathbf{w}}}$ and $\Sigma_{\vec{\mathbf{v}}}^{(t)}$ are diagonal, $\Sigma_{\vec{\mathbf{v}}}^{(t+1)}$ will not be diagonal because the estimator $G(\vec{\mathbf{u}}_n^{(t)})$ operates on the whole vector $\vec{\mathbf{u}}_n^{(t)}$ (but the diagonalization described in Algorithm 3 can be performed successively in every iteration). However, for the diagonalized equivalent model (3.15) and the uncorrelated BG prior (3.17), it can be shown that the V-BAMP iterations preserve the diagonal property of $\Sigma_{\tilde{\vec{\mathbf{v}}}}^{(t)}$ for all t. The proof is sketched in Appendix C. This has the following implications:

- The computation of the estimator $F(\vec{\mathbf{u}}_n; \Sigma_{\vec{\mathbf{v}}})$ and its derivative is significantly simplified.
- The SE becomes B-dimensional instead of $B(B+1)/2$-dimensional. In other words, $B(B+1)/2$ effective noise covariance parameters of $\Sigma_{\vec{\mathbf{v}}}$ are reduced to B effective noise variances, which in turn carry naturally the MSE estimates for each signal vector:

$$\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}(b), \mathbf{x}(b)) = R(\Sigma_{\tilde{\vec{\mathbf{v}}}}^{(t)} - \Sigma_{\tilde{\vec{\mathbf{w}}}})_{b,b}\,.$$

It follows that the transformation has to be done only once after taking the measurements. Since B is typically not large, determining $\mathbf{T}$ is of negligible computational effort.

3.8 Correlated Compressed Sensing

As shown in Section 3.7, an MMV CS measurement can always be transformed into an equivalent one with uncorrelated signal and additive noise vectors. Furthermore, one of the covariance matrices can be chosen to be transformed into the identity matrix. Using these facts, it is possible to analyze the effects of signal correlation on the recovery behavior. In the following, MMV and DCS with the BG prior are considered in order to gain intuition for the effects of signal correlation. Some of the results, however, are not limited to the BG prior.

Noiseless Correlated Compressed Sensing

Assume $B = 2$ jointly sparse measured vectors with $\Sigma_{\vec{\mathbf{x}}} = \left(\begin{smallmatrix} 1 & C \\ C & 1 \end{smallmatrix}\right)$, i.e., the nonzero components have unit variance and the parameter $C \in [-1, 1]$ controls the correlation between the components of $\vec{\mathbf{x}}_n$. In the MMV case, by intuition, as C increases and the (nonzero) components of $\mathbf{x}(1)$ and $\mathbf{x}(2)$ tend to be more similar, one would assume that the second measurement is less informative, since $\mathbf{y}(1)$ and $\mathbf{y}(2)$ tend to be more similar as well. In the limiting case $C = 1$, $\mathbf{x}(1) = \mathbf{x}(2)$ and so $\mathbf{y}(1) = \mathbf{y}(2)$, i.e., the measurements are simply repeated. However, as long as $|C| < 1$, one can apply a whitening (or decorrelation) transform to $\vec{\mathbf{x}}_n$ (e.g., with $\mathbf{T} = (\Sigma_{\vec{\mathbf{x}}})^{-\frac{1}{2}}$), and obtain an equivalent measurement with $\Sigma_{\tilde{\vec{\mathbf{x}}}} = \mathbf{I}_2$. Since the diagonal entries of $\Sigma_{\vec{\mathbf{x}}}$ as well as those of $\Sigma_{\vec{\mathbf{v}}}^{(t)}$ are uneffected by the whitening transform ($\mathbf{T}\Sigma_{\vec{\mathbf{x}}}\mathbf{T}^T$, $\mathbf{T}\Sigma_{\vec{\mathbf{v}}}\mathbf{T}^T$), the resulting MSE is identical to that of the original system.
In contrast, in the DCS the signal correlation does not carry over to the measurement $\vec{\mathbf{y}}_m$. The case of full correlation (i.e., $C = 1$) results in a single measurement vector problem with doubled number of measurements (sampling rate):

$$\Big(\mathbf{y}(1), \mathbf{y}(2)\Big) = \Big(\mathbf{A}(1)\mathbf{x}(1), \mathbf{A}(2)\mathbf{x}(2)\Big)$$
$$\Updownarrow$$
$$\begin{pmatrix} \mathbf{y}(1) \\ \mathbf{y}(2) \end{pmatrix} = \begin{pmatrix} \mathbf{A}(1) \\ \mathbf{A}(2) \end{pmatrix} \mathbf{x}(1)\,.$$

We conclude that the dimensionality/uncertainty of the noiseless MMV measurement is $\mathrm{rank}(\Sigma_{\vec{\mathbf{x}}})$ irrespective of signal correlation, while the noiseless DCS measurement supports the hypothesis that increasing signal correlation is equivalent to an increased sampling rate.

Noisy Correlated Compressed Sensing

Consider the measurement scenario where $\Sigma_{\vec{\mathbf{w}}} = \mathbf{I}_2$ and $\Sigma_{\vec{\mathbf{x}}} = \left(\begin{smallmatrix} 1 & C \\ C & 1 \end{smallmatrix}\right)$, i.e., the nonzero signal components have unit variance and the parameter $C \in [-1, 1]$ controls the correlation between the $B = 2$ components of $\vec{\mathbf{x}}_n$. As C approaches 1, the (nonzero) components in $\mathbf{x}(1)$ and $\mathbf{x}(2)$ become more and more similar. Applying the joint diagonalization transformation in the MMV case one obtains $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_2$ and $\mathrm{diag}(\Sigma_{\tilde{\vec{\mathbf{w}}}}) = (\frac{1}{1+C}, \frac{1}{1-C})^T$. That is, with increasing correlation the measurement can be interpreted as if one of the SNRs tends to 0, while the other tends to double the initial value (or equivalently, one of the noise variances goes to ∞, and the other to $\frac{1}{2}$). In the

limiting case $C = 1$, $\mathbf{x}(1) = \mathbf{x}(2)$, and joint diagonalization with Algorithm 3 can not be performed due to the rank deficiency of $\Sigma_{\tilde{\mathbf{x}}}$ (note that purely joint diagonalization is possible when one does not require $\Sigma_{\tilde{\mathbf{x}}} = \mathbf{I}$). However, the interpretation is still valid: one of the noise variances goes to ∞ and that measurement is uninformative (as the corresponding SNR $\to 0$), thus the MMV measurement turns into a single measurement vector problem with doubled SNR (i.e., half of the original noise variance on the remaining single vector):

$$\begin{aligned}\Big(\mathbf{y}(1), \mathbf{y}(2)\Big) &= \mathbf{A}\Big(\mathbf{x}(1), \mathbf{x}(2)\Big) + \Big(\mathbf{w}(1), \mathbf{w}(2)\Big) \\ &\Updownarrow \\ \mathbf{y}(1) &= \mathbf{A}\mathbf{x}(1) + \mathbf{w}'\,,\end{aligned}$$

where $\mathrm{E}\{\mathsf{w}_n'^{\,2}\} = \frac{1}{2}\,\mathrm{E}\{\mathsf{w}_n(1)^2\}$. The same result can be obtained by the *noise averaging* argument. That is, the signal parts are equal ($\mathbf{A}\mathbf{x}(1) = \mathbf{A}\mathbf{x}(2)$), but observed twice with i.i.d. noise realizations. By combining the observations

$$\frac{\mathbf{y}(1) + \mathbf{y}(2)}{2} = \mathbf{A}\mathbf{x}(1) + \frac{\mathbf{w}(1) + \mathbf{w}(2)}{2}$$

one arrives at a single measurement vector problem with doubled SNR.
In contrast, the DCS scenario turns into a single measurement vector problem with doubled measurement rate and unchanged noise parameters, as the measurement matrices can be stacked:

$$\begin{aligned}\Big(\mathbf{y}(1), \mathbf{y}(2)\Big) &= \Big(\mathbf{A}(1)\mathbf{x}(1), \mathbf{A}(2)\mathbf{x}(2)\Big) + \Big(\mathbf{w}(1), \mathbf{w}(2)\Big) \\ &\Updownarrow \\ \begin{pmatrix}\mathbf{y}(1)\\ \mathbf{y}(2)\end{pmatrix} &= \begin{pmatrix}\mathbf{A}(1)\\ \mathbf{A}(2)\end{pmatrix}\mathbf{x}(1) + \begin{pmatrix}\mathbf{w}(1)\\ \mathbf{w}(2)\end{pmatrix}\,.\end{aligned}$$

We conclude that in the MMV scenario signal correlation amounts to rescaling of the SNRs, while in the the DCS scenario correlation can presumably be interpreted as increasing the sampling rate.

3.9 Replica Analysis

In [83], the replica trick [84] was utilized in order to derive the MSE performance of loopy belief propagation/V-BAMP for the measurement (3.1) and the BG prior

(3.7), assuming $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_B$ and isotropic noise, i.e., $\Sigma_{\vec{\mathbf{w}}} = \sigma_{\mathsf{w}}^2 \mathbf{I}_B$. The derivation is very sophisticated and the generalization to arbitrary signal and noise correlations seems infeasible due to technical difficulties. However, for the MMV scenario, through the joint diagonalization presented in Section 3.7 one is able to circumvent these difficulties, and so it only remains to extend the replica analysis to $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_B$ and an arbitrary diagonal (positive definite) noise covariance matrix $\Sigma_{\vec{\mathbf{w}}} = \mathrm{diag}(\sigma_{\mathsf{w}}^2(1), \ldots, \sigma_{\mathsf{w}}^2(B))$.
In particular, the replica method is capable of predicting the fixed points of loopy belief propagation, which are equivalent to the fixed points of BAMP in the asymptotic regime ($N, M \to \infty$, $R = M/N = \text{const.}$), as a function of the MSE [85, 86]. In [83], the analysis has been extended for V-BAMP, where identical noise variance was assumed and so the overall MSE equals the MSE on each of the B channels. Note that rigorous equivalence between the replica method and SE is not always guaranteed and requires additional technicalities [87]. Following the analysis in [83], we derive an analytical performance prediction for the BAMP for MMV and DCS problems, in which B different additive noise variances and so B different MSE parameters are incorporated.

Consider the signal prior

$$f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = (1-\epsilon)\delta(\vec{\mathbf{x}}_n) + \epsilon \mathcal{N}(\vec{\mathbf{x}}_n; \mathbf{0}, \mathbf{I}_B) \tag{3.18}$$

for $n \in [N]$ and $\vec{\mathbf{x}}_n \in \mathbb{R}^{B\times 1}$. The measurement equations are

$$\mathbf{y}(b) = \mathbf{A}(b)\mathbf{x}(b) + \mathbf{w}(b)\,, \quad b \in [B]\,, \tag{3.19}$$

with $\mathbf{A}(b) \in \mathbb{R}^{M\times N}$ and $\vec{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\vec{\mathbf{w}}})$, where $\Sigma_{\vec{\mathbf{w}}} = \mathrm{diag}(\sigma_{\mathsf{w}}^2(1), \ldots, \sigma_{\mathsf{w}}^2(B))$ is a diagonal matrix carrying the additive noise variances $\sigma_{\mathsf{w}}^2(b)$. We highlight that:

- The analysis is presented for the MMV scenario, i.e., $\mathbf{A}(1) = \ldots = \mathbf{A}(B) = \mathbf{A}$ for the sake of simplicity. The generalization to DCS follows straightforwardly with only more cumbersome notation.
- The analysis assumes that the measurement matrices have normalized rows, but at the end of the derivation the result is translated to the normalized columns case.

The posterior pdf of the estimate $\hat{\mathbf{X}} = (\hat{\mathbf{x}}(1), \ldots, \hat{\mathbf{x}}(B))$ reads

$$f_{\hat{\mathbf{X}}|\mathbf{Y}}(\hat{\mathbf{X}} \mid \mathbf{Y}) = \frac{1}{\mathcal{Z}} \prod_{n=1}^{N} f_{\hat{\vec{\mathbf{x}}}}(\hat{\vec{\mathbf{x}}}_n) \prod_{m=1}^{M} (2\pi|\Sigma_{\vec{\mathbf{w}}}|)^{-\frac{1}{2}} \exp\left(-\left(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}}\right)_m \Sigma_{\vec{\mathbf{w}}}^{-1} \left(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}}\right)_m^T\right), \tag{3.20}$$

where $\mathcal{Z}$ is the partition function

$$\mathcal{Z} = \int_{\mathbb{R}^{N\times B}} \prod_{n=1}^{N} f_{\hat{\vec{\mathbf{x}}}}(\hat{\vec{\mathbf{x}}}_n) \prod_{m=1}^{M} (2\pi|\Sigma_{\vec{\mathbf{w}}}|)^{-\frac{1}{2}} \exp\left(-\left(\mathbf{Y}-\mathbf{A}\hat{\mathbf{X}}\right)_m \Sigma_{\vec{\mathbf{w}}}^{-1} \left(\mathbf{Y}-\mathbf{A}\hat{\mathbf{X}}\right)_m^T\right) \prod_{n=1}^{N} d\hat{\vec{\mathbf{x}}}_n \,.$$

Following the analogy between the measurement model (3.19) and the many-body thermodynamic system [84–86, 88–90] the posterior (3.20) can be interpreted as the Boltzmann measure on a disordered system with Hamiltonian

$$H(\mathbf{X}) = \sum_{n=1}^{N} \log f_{\hat{\vec{\mathbf{x}}}}(\hat{\vec{\mathbf{x}}}_n) + \sum_{m=1}^{M} \left(\mathbf{Y}-\mathbf{A}\hat{\mathbf{X}}\right)_m \Sigma_{\vec{\mathbf{w}}}^{-1} \left(\mathbf{Y}-\mathbf{A}\hat{\mathbf{X}}\right)_m^T . \tag{3.21}$$

The average free energy of the disordered system given by (3.21) characterizes its thermodynamic properties. Evaluating the local extrema in the *free energy function* provides the channel-wise MMSE for the measurement model (3.19) [84–86, 88–90]. Remember that V-BAMP is an approximate MMSE estimator. When the V-BAMP estimate is close to the true MMSE estimate, the replica analysis provides the vector-wise MSE of the estimates at the fixed points of V-BAMP. Assuming self-averaging [84–86, 88–90] the free energy is defined as

$$\mathcal{F} = \lim_{N\to\infty} \frac{1}{N} \mathrm{E}_{\mathbf{A},\mathbf{X},\mathbf{W}} \left\{\log(\mathcal{Z})\right\} . \tag{3.22}$$

In general, this is extremely difficult to evaluate. The replica method [84–86, 88–90] introduces k replicas $\hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^k$ of the estimate $\hat{\mathbf{X}}$ and the free energy (3.22) can be approximated by the *replica trick* [84–86, 90]

$$\mathcal{F} \approx \lim_{N\to\infty} \lim_{k\to 0} \frac{\mathrm{E}_{\mathbf{A},\mathbf{x},\mathbf{w}} \left\{\mathcal{Z}^k\right\} - 1}{Nk} . \tag{3.23}$$

Note that the self-averaging property that leads to (3.22) and the replica trick (3.23), as well as the replica symmetry assumptions are assumed to be valid. Their theoretical justification is, however, still an open problem in mathematical physics [84–86, 88–90]. In order to evaluate the free energy (3.22) via the approximation (3.23), we write

$$\mathrm{E}_{\mathbf{A},\mathbf{X},\mathbf{W}} \left\{\mathcal{Z}^k\right\} = |2\pi\Sigma_{\vec{\mathbf{w}}}|^{-\frac{k}{2}} \mathrm{E}_{\mathbf{X}} \left\{\int_{\mathbb{R}^{N\times B}} \prod_{n=1}^{N} \prod_{a=1}^{k} f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n^a) \prod_{m=1}^{M} \mathbb{X}_m d\vec{\mathbf{x}}_n^a\right\} , \tag{3.24}$$

where

$$\begin{aligned}\mathbb{X}_m &= \mathrm{E}_{\mathbf{A},\mathbf{W}}\left\{\exp\left(-\frac{1}{2}\sum_{a=1}^{k}\vec{\mathbf{v}}_m^a\Sigma_{\vec{\mathbf{w}}}^{-1}(\vec{\mathbf{v}}_m^a)^T\right)\right\}\\ &= \mathrm{E}_{\mathbf{A},\mathbf{W}}\left\{\exp\left(-\frac{1}{2}\sum_{a=1}^{k}\sum_{b=1}^{B}\frac{1}{\sigma_{\mathsf{w}}^2(b)}(v_{m,b}^a)^2\right)\right\}\,,\end{aligned}$$

and

$$\vec{\mathbf{v}}_m^a = (v_{m,1}^a,\ldots,v_{m,B}^a) = \left(\mathbf{A}(\mathbf{X}-\hat{\mathbf{X}}^a)+\mathbf{W}\right)_m\,.$$

The argument of $\mathbb{X}_m$ is rewritten in vector form as

$$\begin{aligned}\mathbb{X}_m &= \mathrm{E}_{\mathbf{A},\mathbf{W}}\left\{-\frac{1}{2}\vec{\mathbf{v}}_m\bar{\Sigma}_{\vec{\mathbf{w}}}^{-1}\vec{\mathbf{v}}_m^T\right\}\\ &= \mathrm{E}_{\mathbf{A},\mathbf{W}}\left\{-\frac{1}{2}\bar{\vec{\mathbf{v}}}_m\bar{\vec{\mathbf{v}}}_m^T\right\}\end{aligned}\tag{3.25}$$

with

$$\vec{\mathbf{v}}_m = (v_{m,1}^1,\ldots,v_{m,1}^k,v_{m,2}^1,\ldots,\ldots,v_{m,B}^k)\in\mathbb{R}^{kB}$$

and

$$\bar{\vec{\mathbf{v}}}_m = \vec{\mathbf{v}}_m\bar{\Sigma}_{\vec{\mathbf{w}}}^{-\frac{1}{2}}$$

with

$$\bar{\Sigma}_{\vec{\mathbf{w}}} = \Sigma_{\vec{\mathbf{w}}}\otimes\mathbf{I}_k\,.$$

Let us evaluate the covariance matrix $\mathbf{G}_m = \mathrm{Cov}\{\bar{\vec{\mathbf{v}}}_m\}$. It is composed of $B\times B$ blocks of size $k\times k$:

1. The main diagonal of $\mathbf{G}_m$ consists of entries $g_1(b) = \mathrm{E}_{\mathbf{A},\mathbf{w}}\{\frac{1}{\sigma_{\mathsf{w}}^2(b)}(v_{m,b}^a)^2\}$, which is different in each of the B blocks but identical within a block.

2. The remaining entries in the blocks of the main diagonal are $g_2(b) = \mathrm{E}_{\mathbf{A},\mathbf{w}}\{\frac{1}{\sigma_{\mathsf{w}}^2(b)}v_{m,b}^a v_{m,b}^{a'}\}$, which are different in each block but identical within a block.

3. The diagonal entries of the off-diagonal blocks are $g_3(b,b') = \mathrm{E}_{\mathbf{A},\mathbf{w}}\{v_{m,b}^a v_{m,b'}^a\}$.

4. The off-diagonal entries of the off-diagonal blocks are $g_4(b,b') = \mathrm{E}_{\mathbf{A},\mathbf{w}}\{v_{m,b}^a v_{m,b'}^{a'}\}$.

For random measurement matrices $\mathbf{A}$ (e.g., Bernoulli or Gaussian, see Section 1.1.4), and due to (i) $\vec{\mathbf{x}}_n^a$ following the same distribution as $\vec{\mathbf{x}}_n$, (ii) the replica symmetry

[85, 86], these values turn out to be

$$
\begin{aligned}
g_1(b) &= \frac{1}{\sigma_{\mathsf{w}}^2(b)} \, \mathrm{E}_{\mathbf{A},\mathbf{w}}\{(v_{m,b}^a)^2\} \\
&= \frac{1}{\sigma_{\mathsf{w}}^2(b)} \Big(\frac{1}{N} \sum_n^N (x_n(b) - \hat{x}_n^a(b))^2 + \sigma_{\mathsf{w}}^2(b) \Big) \,, \\
g_2(b) &= \frac{1}{\sigma_{\mathsf{w}}^2(b)} \, \mathrm{E}_{\mathbf{A},\mathbf{w}}\{v_{m,b}^a v_{m,b}^{a'}\} \\
&= \frac{1}{\sigma_{\mathsf{w}}^2(b)} \Big(\frac{1}{N} \sum_n^N (x_n(b) - \hat{x}_n^a(b))(x_n(b) - \hat{x}_n^{a'}(b)) + \sigma_{\mathsf{w}}^2(b) \Big) \,, \\
g_3(b,b') &= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')} \, \mathrm{E}_{\mathbf{A},\mathbf{w}}\{v_{m,b}^a v_{m,b'}^a\} \\
&= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')} \Big(\frac{1}{N} \sum_n^N (x_n(b) - \hat{x}_n^a(b))(x_n(b') - \hat{x}_n^a(b)) \Big) \,, \\
g_4(b,b') &= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')} \, \mathrm{E}_{\mathbf{A},\mathbf{w}}\{v_{m,b}^a v_{m,b'}^{a'}\} \\
&= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')} \Big(\frac{1}{N} \sum_n^N (x_n(b) - \hat{x}_n^a(b'))(x_n(b) - \hat{x}_n^{a'}(b')) \Big) \,.
\end{aligned}
$$

By introducing the auxiliary quantities

$$
\begin{aligned}
m_a(b,b') &= \frac{1}{N} \sum_{n=1}^N \hat{x}_n^a(b) x_n(b') \,, \\
Q_a(b,b') &= \frac{1}{N} \sum_{n=1}^N \hat{x}_n^a(b) \hat{x}_n^a(b') \,, \\
q_{aa'}(b,b') &= \frac{1}{N} \sum_{n=1}^N \hat{x}_n^a(b) \hat{x}_n^{a'}(b') \,, \\
q_0(b,b') &= \frac{1}{N} \sum_{n=1}^N x_n(b) x_n(b') \,,
\end{aligned}
$$

the covariance values can be written as

$$
\begin{aligned}
g_1(b) &= \frac{1}{\sigma_{\mathsf{w}}^2(b)}(\epsilon - 2m_a(b,b) + Q_a(b,b) + \sigma_w^2(b))\,,\\
g_2(b) &= \frac{1}{\sigma_{\mathsf{w}}^2(b)}(\epsilon - (m_a(b,b) + m_{a'}(b,b)) + q_{aa'}(b,b) + \sigma_w^2(b))\,,\\
g_3(b,b') &= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')}(q_0(b,b') - (m_a(b,b) + m_{a'}(b,b)) + q_{aa'}(b,b))\,,\\
g_4(b,b') &= \frac{1}{\sigma_{\mathsf{w}}(b)\sigma_{\mathsf{w}}(b')}(q_0(b,b) - (m_a(b,b') + m_{a'}(b',b)) + q_{aa'}(b',b'))\,.
\end{aligned}
$$

The pdf of $\bar{\vec{\mathbf{v}}}_m$ is approximated using the central limit theorem by a multivariate Gaussian distribution as

$$
f_{\bar{\vec{\mathbf{v}}}_m}\left(\bar{\vec{\mathbf{v}}}_m\right) = \mathcal{N}\left(\bar{\vec{\mathbf{v}}}_m; \mathbf{0}, \mathbf{G}_m\right)\,. \tag{3.26}
$$

Combining (3.25) and (3.26) one obtains

$$
\begin{aligned}
\mathbb{X}_m &= \mathrm{E}_{\bar{\vec{\mathbf{v}}}_m}\left\{\exp\left(-\frac{1}{2}\bar{\vec{\mathbf{v}}}_m\bar{\vec{\mathbf{v}}}_m^T\right)\right\}\\
&= \int_{\mathbb{R}^{kB}} \exp\left(-\frac{1}{2}\bar{\vec{\mathbf{v}}}_m\bar{\vec{\mathbf{v}}}_m^T\right)\mathcal{N}\left(\bar{\vec{\mathbf{v}}}_m; \mathbf{0}, \mathbf{G}_m\right) d\bar{\vec{\mathbf{v}}}_m\\
&= \int_{\mathbb{R}^{kB}} |2\pi\mathbf{G}_m|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\bar{\vec{\mathbf{v}}}_m(\mathbf{I}_{kB} + \mathbf{G}_m^{-1})\bar{\vec{\mathbf{v}}}_m^T\right) d\bar{\vec{\mathbf{v}}}_m\\
&= |2\pi\mathbf{G}_m|^{-\frac{k}{2}}|2\pi(\mathbf{G}_m^{-1} + \mathbf{I}_{kB})^{-1}|^{\frac{k}{2}} \int_{\mathbb{R}^{kB}} \mathcal{N}\left(\bar{\vec{\mathbf{v}}}_m; \mathbf{0}, (\mathbf{I}_{kB} + \mathbf{G}_m^{-1})^{-1}\right) d\bar{\vec{\mathbf{v}}}_m\\
&= |\mathbf{I}_{kB} + \mathbf{G}_m|^{-\frac{1}{2}}\,.
\end{aligned}
$$

In the Bayesian setting the distribution of $\vec{\mathbf{x}}_n$ matches the distribution of $\hat{\bar{\vec{\mathbf{x}}}}_n$ and that of the replicas $\hat{\bar{\vec{\mathbf{x}}}}_n^a$, thus $g_3 = g_4 = 0$. Furthermore, due to the replica symmetry [85, 86] $m_a(b,b) = m_{a'}(b,b) = m$, $Q_a(b,b) = Q(b)$, and $q_{aa'}(b,b) = q(b)$. The covariance $\mathbf{G}_m$ is a structured matrix that, due to its block structure, can easily be constructed using all-ones matrices, identity matrices, and Kronecker products. Its kB eigenvalues turn out to be:

$$
\begin{aligned}
\alpha_1(b) &= g_1(b) + (k-1)g_2(b)\,, && \times 1\,, && b \in [B]\,,\\
\alpha_2(b) &= g_1(b) - g_2(b)\,, && \times (k-1)\,, && b \in [B]\,,
\end{aligned}
$$

with $\alpha_1(b)$ having multiplicity 1 ($b \in [B]$) and $\alpha_2(b)$ having multiplicity $k-1$ ($b \in [B]$). Thus,

$$
\begin{aligned}
|\mathbf{I}_{kB} + \mathbf{G}_m|^{-\frac{1}{2}} &= \left(\prod_{b=1}^{B}(1+\alpha_1(b))(1+\alpha_2(b))^{k-1}\right)^{-\frac{1}{2}} \\
&= \left[\prod_{b=1}^{B}\left(1 + k\frac{\epsilon - 2m(b) + q(b) + \sigma_{\mathsf{w}}^2(b)}{\sigma_{\mathsf{w}}^2(b) + Q(b) - q(b)}\right)\prod_{b=1}^{B}\left(1 + \frac{1}{\sigma_{\mathsf{w}}^2(b)}(Q(b) - q(b))\right)^{k-1}\right]^{-\frac{1}{2}} .
\end{aligned}
$$

Using the Taylor series approximation

$$
\exp(x) \approx 1 + x \Rightarrow (1+x)^{-\frac{1}{2}} \approx \exp\left(-\frac{x}{2}\right) ,
$$

one arrives at

$$
\lim_{k\to 0} \mathbb{X}_m \approx \exp\left(-\frac{k}{2}\sum_{b=1}^{B}\frac{\epsilon - 2m(b) + q(b) + \sigma_{\mathsf{w}}^2(b)}{\sigma_{\mathsf{w}}^2(b) + Q(b) - q(b)} - \log\left(Q(b) - q(b) + \sigma_{\mathsf{w}}^2(b)\right) + \log\left(\sigma_{\mathsf{w}}^2(b)\right)\right) .
$$

Following the derivation in [83, App.], (3.24) can be written as

$$
\mathrm{E}_{\mathbf{A},\mathbf{X},\mathbf{W}}\left\{\mathcal{Z}^k\right\} = \int \exp\left(kN\Phi(m, \hat{m}, q, \hat{q}, Q, \hat{Q})\right) dm\, d\hat{m}\, dq\, d\hat{q}\, dQ\, d\hat{Q} .
$$

Remember that one is only interested in the stationary points of the free energy expression (3.24) [83]. Thus, we set

$$
\begin{aligned}
\mathcal{F} &= \Phi(\{m(b)^*, \hat{m}(b)^*, q(b)^*, \hat{q}(b)^*, Q(b)^*, \hat{Q}(b)^*\}_{b=1,\ldots,B}) \\
&= \frac{1}{2}\sum_{b=1}^{B}(Q(b)\hat{Q}(b) - 2m(b)\hat{m}(b) + q(b)\hat{q}(b)) - \frac{R}{2}\log\left(|2\pi\Sigma_{\vec{\mathbf{w}}}|\right) \\
&\quad - \frac{R}{2}\sum_{b=1}^{B}\left(\frac{\epsilon - 2m(b) + q(b) + \sigma_{\mathsf{w}}^2(b)}{Q(b) - q(b) + \sigma_{\mathsf{w}}^2(b)} + \log\left(Q(b) - q(b) + \sigma_{\mathsf{w}}^2(b)\right) - \log\left(\sigma_{\mathsf{w}}^2(b)\right)\right) \\
&\quad + \int_{\mathbb{R}^B} f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}) \int_{\mathbb{R}^B} \log \int_{\mathbb{R}^B} f_{\hat{\vec{\mathbf{x}}}}(\hat{\vec{\mathbf{x}}}) \\
&\quad \prod_{b=1}^{B} \exp\Big(-\frac{1}{2}\hat{q}(b)\,\hat{x}(b)^2 + \hat{m}(b)\,\hat{x}(b)x(b) + \sqrt{\hat{m}(b)}\,\hat{x}(b)h(b)\Big) d\hat{\vec{\mathbf{x}}}\,\mathcal{D}\vec{\mathbf{h}}\,d\vec{\mathbf{x}} ,
\end{aligned}
\tag{3.27}
$$

where * denotes stationary points, and the second integration is over a Gaussian measure, i.e., $\mathcal{D}\mathbf{h} = \mathcal{N}(\vec{\mathbf{h}}; \mathbf{0}, \mathbf{I}_B)d\vec{\mathbf{h}} = \prod_{b=1}^{B}\mathcal{N}(h(b); 0, 1)dh(b)$. The stationary points

are obtained by differentiation as

$$\frac{d\Phi}{dm(b)} = 0 \Rightarrow \hat{m}(b)^* = \frac{R}{E(b) + \sigma_{\mathsf{w}}^2(b)} = \gamma(b)\,,$$
$$\frac{d\Phi}{dq(b)} = 0 \Rightarrow \hat{q}(b)^* = \frac{R}{E(b) + \sigma_{\mathsf{w}}^2(b)} = \gamma(b)\,,$$
$$\frac{d\Phi}{dQ(b)} = 0 \Rightarrow \hat{Q}(b)^* = 0\,,$$

where we used the substitution[1] $E(b) = Q(b) - q(b)$, and that in the Bayesian setting $q(b)^* = m(b)^*$, and $Q(b)^* = \epsilon$. As $N \to \infty$, $E(b) = \mathrm{MSE}(\hat{\mathbf{x}}(b), \mathbf{x}(b))$. Substituting back into (3.27) and using $\vec{\mathbf{E}} = (E(1), \ldots, E(B))^T$ one obtains

$$\begin{aligned}\mathcal{F}(\vec{\mathbf{E}}, \Sigma_{\vec{\mathbf{w}}}) = &-\frac{R}{2}\sum_{b=1}^{B}\left(\log\Big(2\pi(\sigma_{\mathsf{w}}^2(b) + E(b))\Big) + \frac{\epsilon + \sigma_{\mathsf{w}}^2(b)}{E(b) + \sigma_{\mathsf{w}}^2(b)}\right)\\ &+ \int_{\mathbb{R}^B} f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}) \int_{\mathbb{R}^B} \log\Bigg(\int_{\mathbb{R}^B} f_{\hat{\vec{\mathbf{x}}}}(\hat{\vec{\mathbf{x}}})\\ &\prod_{b=1}^{B} \exp\Big(-\frac{1}{2}\gamma(b)\,\hat{x}(b)^2 + \gamma(b)\hat{x}(b)x(b) + \sqrt{\gamma(b)}\hat{x}(b)h(b)\Big)d\hat{\vec{\mathbf{x}}}\Bigg)\,\mathcal{D}\vec{\mathbf{h}}\,d\vec{\mathbf{x}}\,.\end{aligned}$$

Inserting the signal prior (3.18) results in

$$\begin{aligned}\mathcal{F}(\vec{\mathbf{E}}, \Sigma_{\vec{\mathbf{w}}}) = &-\frac{R}{2}\sum_{b=1}^{B}\left(\log\Big(2\pi(\sigma_{\mathsf{w}}^2(b) + E(b))\Big) + \frac{\epsilon + \sigma_{\mathsf{w}}^2(b)}{E(b) + \sigma_{\mathsf{w}}^2(b)}\right)\\ &+ (1-\epsilon)\int \log\Big((1-\epsilon) + \epsilon\int \exp\Big(-\frac{1}{2}\gamma(b)\hat{x}^2 + \sqrt{\gamma(b)}\hat{x}(b)h(b)\Big)\mathcal{D}\mathbf{x}\Big)\,\mathcal{D}\vec{\mathbf{h}}\\ &+ \epsilon\int\int \log\Bigg((1-\epsilon) +\\ &\epsilon\int \exp\Big(-\frac{1}{2}\gamma(b)\hat{x}(b)^2 + \gamma(b)\hat{x}(b)x(b) + \sqrt{\gamma(b)}\hat{x}(b)h(b)\Big)\mathcal{D}\hat{\vec{\mathbf{x}}}\Bigg)\mathcal{D}\vec{\mathbf{h}}\,\mathcal{D}\vec{\mathbf{x}}\,,\end{aligned}$$

[1]Following the proof: $\frac{1}{N}\sum_n \hat{x}_n^2 - x_n^2 = \frac{1}{N}\sum_n (x_n + e_n)^2 - x_n^2 = \frac{1}{N}\sum_n x_n^2 + 2x_n e_n + e_n^2 - x_n^2 = \frac{1}{N}\sum_n e_n^2 = \frac{1}{N}\sum_n (\hat{x}_n - x_n)^2 \to \mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x})$, where we used that the estimation error e_n and x_n are uncorrelated and that $N \to \inf$.

with the measures $\mathcal{D}\vec{\mathbf{x}}$ and $\mathcal{D}\hat{\vec{\mathbf{x}}}$ analogously to $\mathcal{D}\vec{\mathbf{h}}$. Writing out the integration measures and further simplifying leads to

$$\begin{aligned}\mathcal{F}(\vec{\mathbf{E}}, \Sigma_{\vec{\mathbf{w}}}) = &-\frac{R}{2}\sum_{b=1}^{B}\left(\log\left(2\pi(\sigma_{\mathsf{w}}^2(b)+E(b))\right)+\frac{\epsilon+\sigma_{\mathsf{w}}^2(b)}{E(b)+\sigma_{\mathsf{w}}^2(b)}-\frac{\gamma(b)(1+\epsilon\gamma(b))}{R(1+\gamma(b))}\right)\\ &+\int\log\left(\epsilon\prod_{b=1}^{B}(1+\gamma(b))^{-\frac{1}{2}}+(1-\epsilon)\exp\Big(-\frac{1}{2}\sum_{b=1}^{B}\gamma(b)h^2(b)\Big)\right)\mathcal{D}\vec{\mathbf{h}}\\ &+\int\log\left(\epsilon\prod_{b=1}^{B}(1+\gamma(b))^{-\frac{1}{2}}+(1-\epsilon)\exp\Big(-\frac{1}{2}\sum_{b=1}^{B}\frac{\gamma(b)}{1+\gamma(b)}h^2(b)\Big)\right)\mathcal{D}\vec{\mathbf{h}}\,.\end{aligned}$$

We underline that the result coincides with [83, eq.(14)] when the noise variances are identical, i.e., $\sigma_{\mathsf{w}}^2(b)=\sigma_{\mathsf{w}}^2$, $\forall b\in[B]$.

Free energy function rescaling

Remember that the above derivation assumes normalized rows in $\mathbf{A}(b)$, $b\in[B]$. In order to arrive at a free energy function that is valid for measurement matrices with normalized columns we rescale the measurement equation and replace $\sigma_{\mathsf{w}}^2(b)$ with $R\sigma_{\mathsf{w}}^2(b)$:

$$\begin{aligned}\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{w} \quad\Leftrightarrow\quad \frac{1}{\sqrt{R}}\mathbf{y}&=\bar{\mathbf{A}}\mathbf{x}+\frac{1}{\sqrt{R}}\mathbf{w}\\ &=\bar{\mathbf{y}}=\bar{\mathbf{A}}\mathbf{x}+\bar{\mathbf{w}}\,,\end{aligned}$$

where $\bar{\mathbf{A}}$ has normalized columns and $\bar{\mathsf{w}}_m\sim\mathcal{N}(0,\frac{\sigma_{\mathsf{w}}^2}{R})$ if $\mathsf{w}_m\sim\mathcal{N}(0,\sigma_{\mathsf{w}}^2)$. The free energy function given that the measurement matrices have normalized columns becomes

$$\begin{aligned}\mathcal{F}(\vec{\mathbf{E}}, \Sigma_{\vec{\mathbf{w}}}) = &-\frac{R}{2}\sum_{b=1}^{B}\left(\log\left(2\pi(R\sigma_{\mathsf{w}}^2(b)+E(b))\right)+\frac{\epsilon+R\sigma_{\mathsf{w}}^2(b)}{E(b)+R\sigma_{\mathsf{w}}^2(b)}-\frac{\gamma(b)(1+\epsilon\gamma(b))}{R(1+\gamma(b))}\right)\\ &+\int\log\left(\epsilon\prod_{b=1}^{B}(1+\gamma(b))^{-\frac{1}{2}}+(1-\epsilon)\exp\Big(-\frac{1}{2}\sum_{b=1}^{B}\gamma(b)h^2(b)\Big)\right)\mathcal{D}\vec{\mathbf{h}}\\ &+\int\log\left(\epsilon\prod_{b=1}^{B}(1+\gamma(b))^{-\frac{1}{2}}+(1-\epsilon)\exp\Big(-\frac{1}{2}\sum_{b=1}^{B}\frac{\gamma(b)}{1+\gamma(b)}h^2(b)\Big)\right)\mathcal{D}\vec{\mathbf{h}}\end{aligned} \tag{3.28}$$

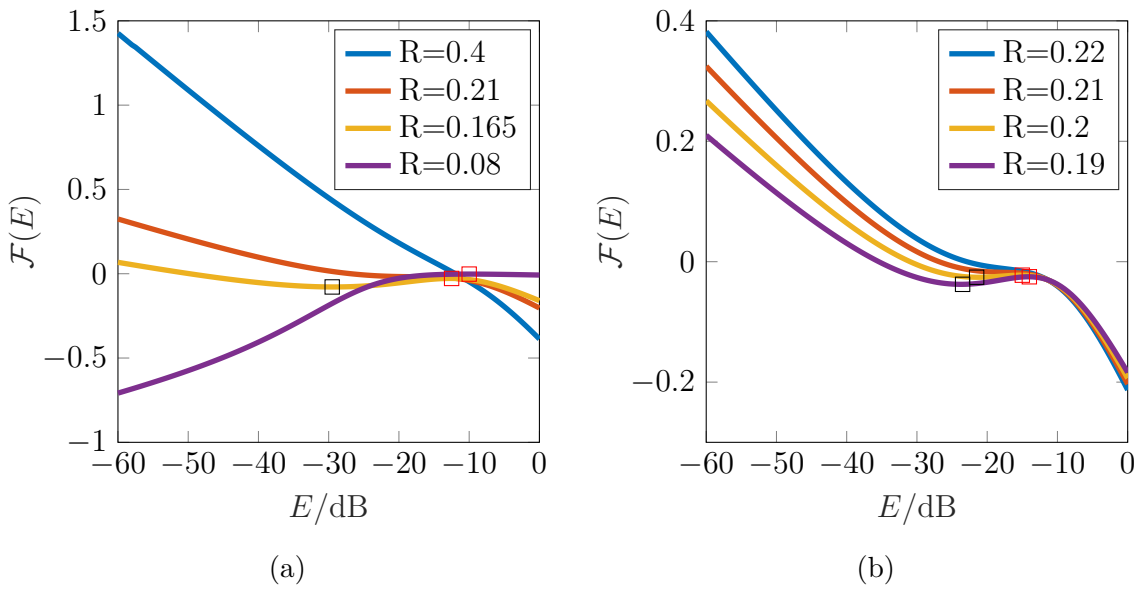
with

$$\gamma(b)=\frac{R}{E(b)+R\sigma_{\mathsf{w}}^2(b)}\,.$$

Discussion

The free energy function (3.28) cannot be further simplified, because the arguments of the exponents are weighted sums of χ^2-distributed random variables (for which a closed form pdf is not known), and the antiderivative of $1 + \log(x)$ is not known as well. Thus, it remains to numerically evaluate the integrals and $\mathcal{F}(\vec{\mathbf{E}})$ in the region of interest on a high resolution grid. In Figure 3.6, free energy functions for different rates and $B = 1$, $\epsilon = 0.1$, $\sigma_{\mathsf{x}}^2 = 1$, and no additive noise are shown. The values of $E = E(1)$ at the local extrema of $\mathcal{F}(E)$ correspond to fixed points of (V)BAMP (i.e., where $\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} = \hat{\mathbf{x}}$) in terms of the $\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}, \mathbf{x})$, and the local maxima to stable fixed points. Starting from a large MSE (i.e., *from the right* on the E-axis), V-BAMP typically converges to the local maximum with the larger MSE, whereas the MMSE of the CS measurement is the smallest E for which $\mathcal{F}(E)$ is a local maximum. Comparing with Figure 2.5 (Section 2.3.1) one can establish a match:

- At $R = 0.4 > R_{\mathrm{PT}}$ there is one local maximum at $E = 0$.
- At $R = 0.165 < R_{\mathrm{PT}}$ there are two local maxima, one at $E \approx -10\mathrm{dB}$, and one at $E = 0$. The local minimum is at $E \approx -30\mathrm{dB}$, corresponding to the second crossing in Figure 2.5c.
- At $R = 0.08 < R_{\mathrm{PT}}$ there are two local maxima, one at $E \approx -10\mathrm{dB}$ (corresponding to the crossing in 2.5d), and one in the limiting case at $E \to 0$. In particular, as $R = 0.08$ lies below the second order PT, the two local maxima are not separated by a local minimum.
- The R_{PT} lies at $R \approx 0.21$, above which V-BAMP will converge to $\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = 0$ whp (successful), and below which V-BAMP will converge to a relatively high MSE whp (unsuccessful). This is visualized in Figure 3.3b, where the local minimum and the first local maximum simultaneously (dis)appear at $R \approx 0.21$, and the only remaining maximum is at $E = 0$.



(a) (b)

Fig. 3.6 1-D free energy functions for different rates ($\epsilon = 0.1$ and $\sigma_{\mathsf{w}}^2 = 0$).

Next, the noisy uniform case is investigated. In Figure 3.7, free energy functions are shown for different rates, $B = 1$, $\epsilon = 0.1$, and additive noise with $\sigma_{\mathsf{w}}^2 = -35$dB. Comparing with Figure 2.6 (Section 2.3.1), one can establish a match:

- At $R = 0.4 > R_{\text{PT}}$ and $R = 0.21 \gtrsim R_{\text{PT}}$ there is a local maximum at $E \approx -32$dB and $E \approx -30$dB, corresponding to the crossing on Figure 2.6a and 2.6b, respectively.

- At $R = 0.165 < R_{\text{PT}}$ and $R = 0.08 < R_{\text{PT}}$, there is a local maximum at $E = -12$dB and $E = -10$dB, corresponding to the crossing on Figure 2.6c and 2.6d, respectively.

- The R_{PT} lies at $R \approx 0.21$, above which V-BAMP will converge to $\text{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = 0$ whp (successful),and below which V-BAMP will converge to a high $\text{MSE}(\hat{\mathbf{x}}, \mathbf{x})$ whp (unsuccessful). This is visualized in Figure 3.7(b), where a second local maximum and a local minimum appear at $R \approx 0.17$, and disappear at $R \approx 0.21$. That is, between $R = 0.17$ and $R = 0.21$, the MMSE of the CS measurement (as predicted by the replica analysis) is typically not reached by V-BAMP.

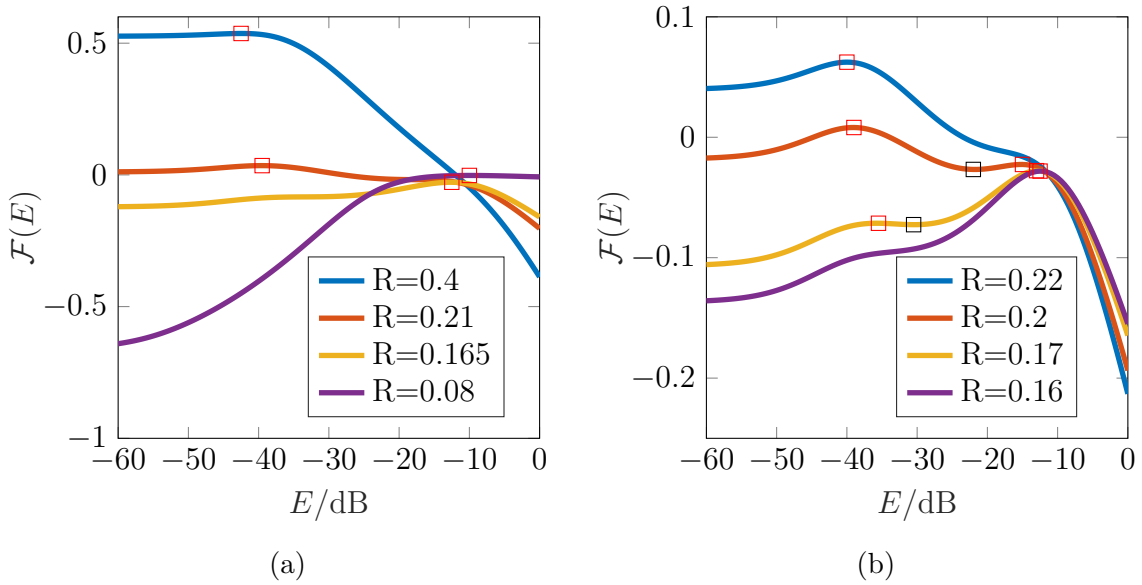


Fig. 3.7 1-D free energy functions for different rates ($\epsilon = 0.1$ and $\sigma_{\mathsf{w}}^2 = -35$dB).

In order to understand the connection between the free energy function and SE in full depth, in Figure 3.8 and Figure 3.9 the free energy functions are shown directly below the 2-dimensional SE with matched axes. The parameters are again $B = 2$, $\epsilon = 0.1$, $\Sigma_{\vec{\mathsf{x}}} = \mathbf{I}_2$, and $(\Sigma_{\vec{\mathsf{w}}})_{b,b} = -35$dB. Let us use $\mathrm{MSE}^{(t)}(b)$ for $\mathrm{MSE}(\hat{\mathbf{x}}^{(t)}(b), \mathbf{x}(b))$, and $\mathrm{MSE}(b)$ for $\mathrm{MSE}^{(t)}(b)$ as $t \to \infty$. In the upper two rows, the axes are the MSE values of the vectors $b = 1$ respectively $b = 2$, in dB. In the first row, the arrows point from one pair of MSE $(\mathrm{MSE}^{(t)}(1), \mathrm{MSE}^{(t)}(2))$ towards the MSE pair $(\mathrm{MSE}^{(t+1)}(1), \mathrm{MSE}^{(t+1)}(2))$ predicted by the SE equation (note that the arrows are scaled for clarity). In the second row the *streamlines* corresponding to the arrows in the first row are plotted, i.e., the curves whose tangents are the arrows. In the third row, the free energy $\mathcal{F}(\vec{\mathbf{E}}) = \mathcal{F}(E(1), E(2))$ is shown, where the brightness of the shades corresponds to the value of the free energy function (i.e., brighter means larger value). The blue arrows depict the gradient vectors of $\mathcal{F}(\vec{\mathbf{E}})$, while the black lines are isolines. In all three plots, stable fixed points (of the SE equation)/sinks (of the streamlines)/local maxima (of the free energy function) are denoted by a red square, while unstable fixed points/sources/saddle points are denoted by a black square. Observe that at a low rate (e.g., $R = 0.11$) there is only one stable fixed point/local maximum, at a pair of relatively high MSE $/\vec{\mathbf{E}}$. As the measurement rate increases, a second stable fixed point/local maximum appears at a pair of lower MSE $/\vec{\mathbf{E}}$ (which is the component-wise MMSE), together with an unstable fixed point/saddle point. As the rate increases further, the value of the second local maximum rises, while the fixed

point/saddle point translates towards the first fixed point/local maximum. At an even higher rate the unstable fixed point/saddle point merges with the first stable fixed point/local maximum and they annihilate each other, leaving only the stable fixed point/local maximum with the pair of lower MSE $/\vec{\mathbf{E}}$. We conclude that the sampling rate region in which V-BAMP does not reach the MMSE performance is where two local maxima of $\mathcal{F}(\vec{\mathbf{E}})$ simultaneously exist, as V-BAMP typically converges to the fixed point with the higher component-wise MSE, while the component-wise MMSE corresponds to the fixed point with the lower component-wise MSE. Comparing the arrows of the SE prediction (first row) and the gradient vectors of the free energy, one can establish a match and interpret V-BAMP empirically as a gradient ascend on the free energy function, which is initialized at a high component-wise MSE.

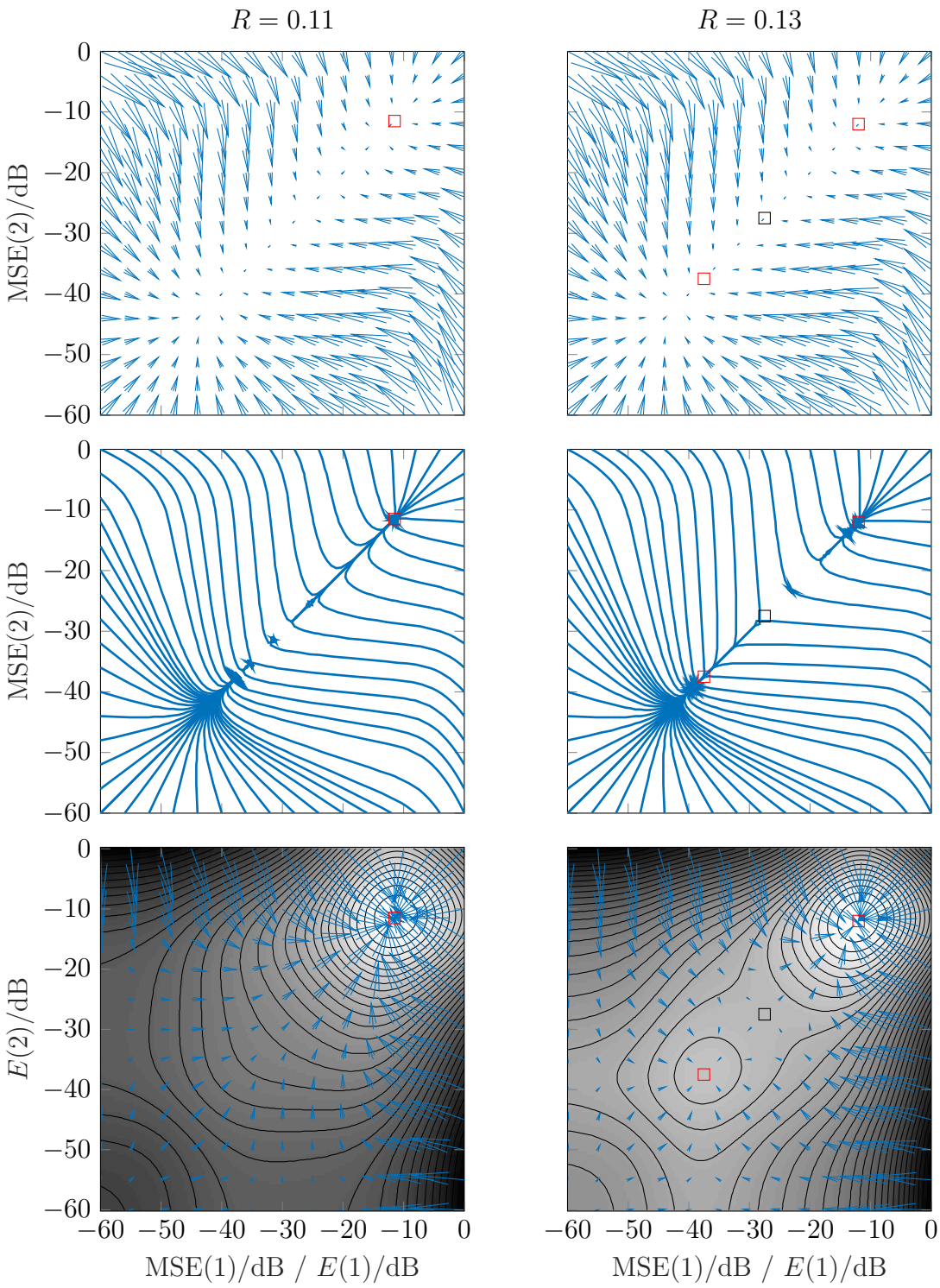


Fig. 3.8 The 2-dimensional (symmetric) SE and replica analysis MSE prediction ($\Sigma_{\vec{\mathsf{x}}} = \mathbf{I}_2$, $\sigma_{\mathsf{w}}^2(1) = \sigma_{\mathsf{w}}^2(2) = -35$dB). Top row: the SE prediction on the 2-D MSE plane. Middle row: streamline curves whose tangents are the arrows of the top row. Bottom row: 2-D symmetric free energy function with isolines and gradient vectors: the brightness of the shade represents the free energy value (i.e., brighter means larger value). Red/black squares are placed at stable/unstable fixed points and local maxima/saddle points, respectively.

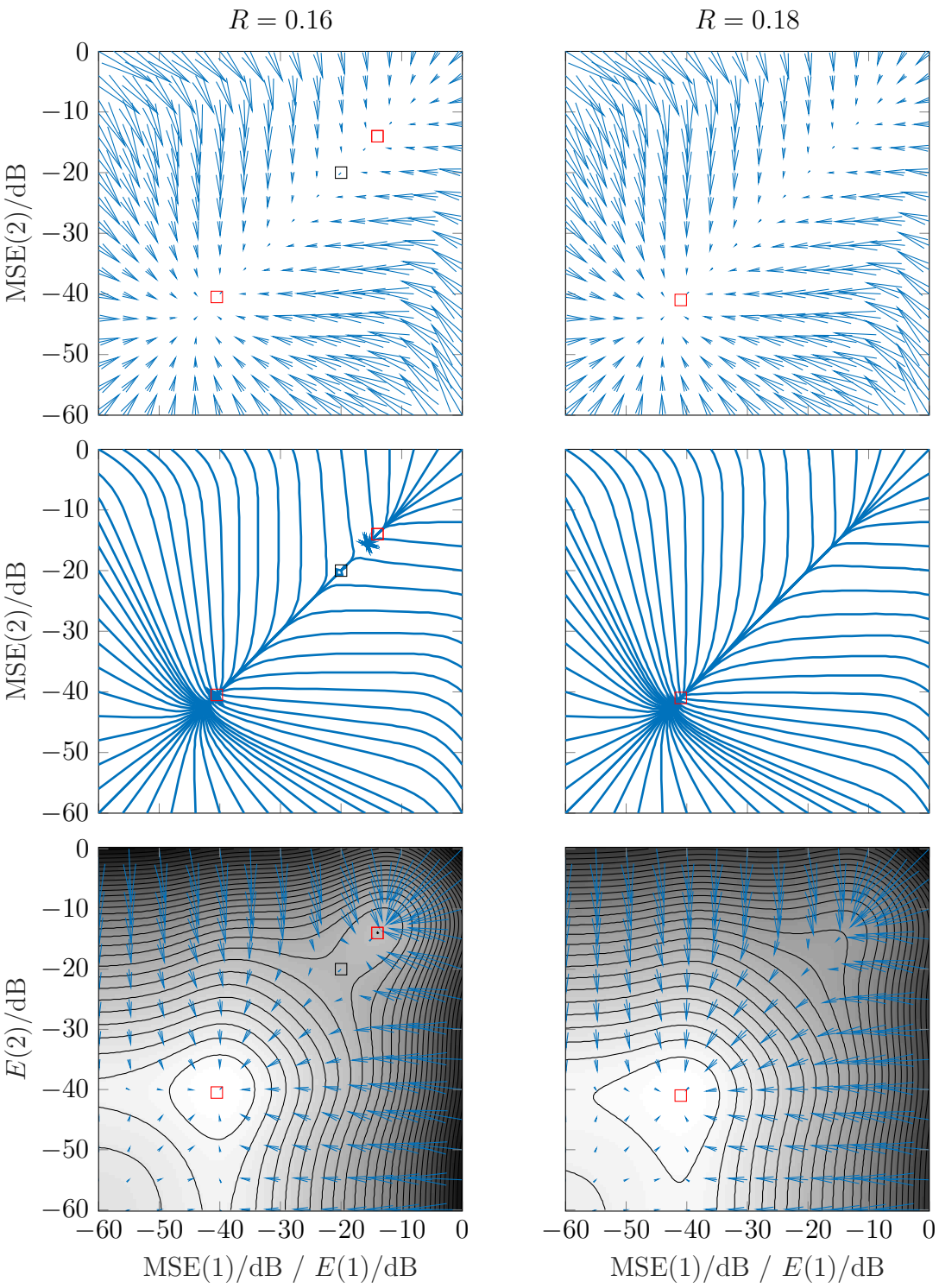


Fig. 3.9 The 2-dimensional (symmetric) SE and replica analysis MSE prediction ($\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_2$, $\sigma_{\mathsf{w}}^2(1) = \sigma_{\mathsf{w}}^2(2) = -35$dB). Top row: the SE prediction on the 2-D MSE plane. Middle row: streamline curves whose tangents are the arrows of the top row. Bottom row: 2-D symmetric free energy function with isolines and gradient vectors: the brightness of the shade represents the free energy value (i.e., brighter means larger value). Red/black squares are placed at stable/unstable fixed points and local maxima/saddle points, respectively.

The non-uniform noise case

In Figure 3.10 and Figure 3.11, the SE and replica analysis prediction are compared, this time for the anisotropic noise case with $\epsilon = 0.1$, $\sigma_{\mathsf{w}}^2(1) = -45\text{dB}$, and $\sigma_{\mathsf{w}}^2(2) = -25\text{dB}$. Analogously to the uniform case, one observes phase transitions as the rate R increases: at a low rate one stable fixed point/local maximum (at a pair of high $(\text{MSE}(1), \text{MSE}(2))/\vec{\mathbf{E}}$ pair) is present; as the rate increases, a second stable fixed point/local maximum appears simultaneously with an unstable fixed point/saddle point; at an even higher rate the unstable fixed point/saddle point merges into the first stable fixed point/local maximum and only one stable fixed point/local maximum at a pair of low $(\text{MSE}(1), \text{MSE}(2))/\vec{\mathbf{E}}$ pair remains. The match between the SE and the replica analysis prediction confirms that the generalization to arbitrary diagonal noise covariance matrices $\Sigma_{\vec{\mathsf{w}}}$ is meaningful.

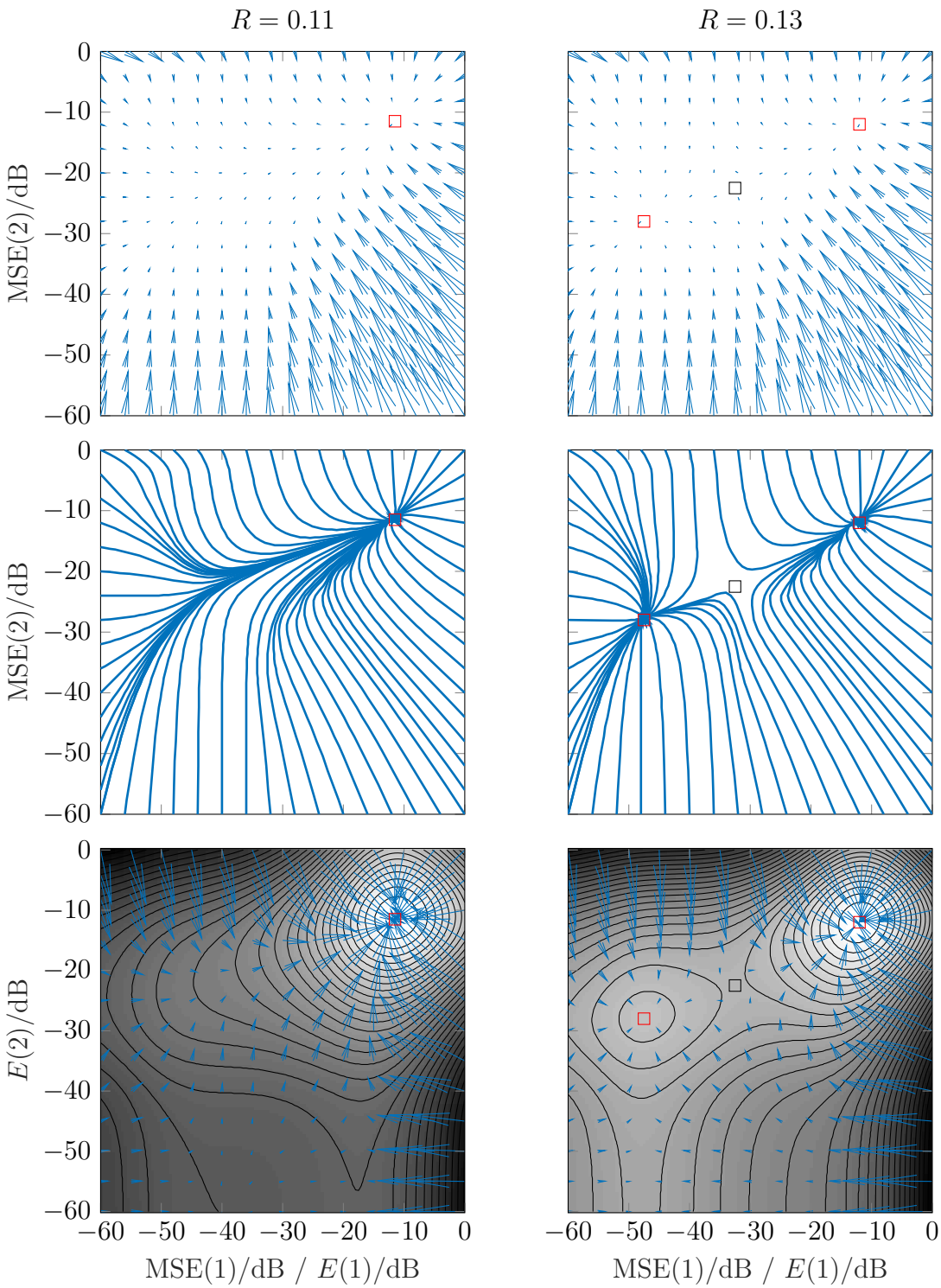


Fig. 3.10 The 2-dimensional (asymmetric) SE and replica analysis MSE prediction ($\Sigma_{\tilde{\mathbf{x}}} = \mathbf{I}_2$, $\sigma^2_{\mathsf{w}}(1) = -45$dB, $\sigma^2_{\mathsf{w}}(2) = -25$dB). Top row: the SE prediction on the 2-D MSE plane. Middle row: streamline curves whose tangents are the arrows of the top row. Bottom row: 2-D symmetric free energy function with isolines and gradient vectors: the brightness of the shade represents the free energy value (i.e., brighter means larger value). Red/black squares are placed at stable/unstable fixed points and local maxima/saddle points, respectively.

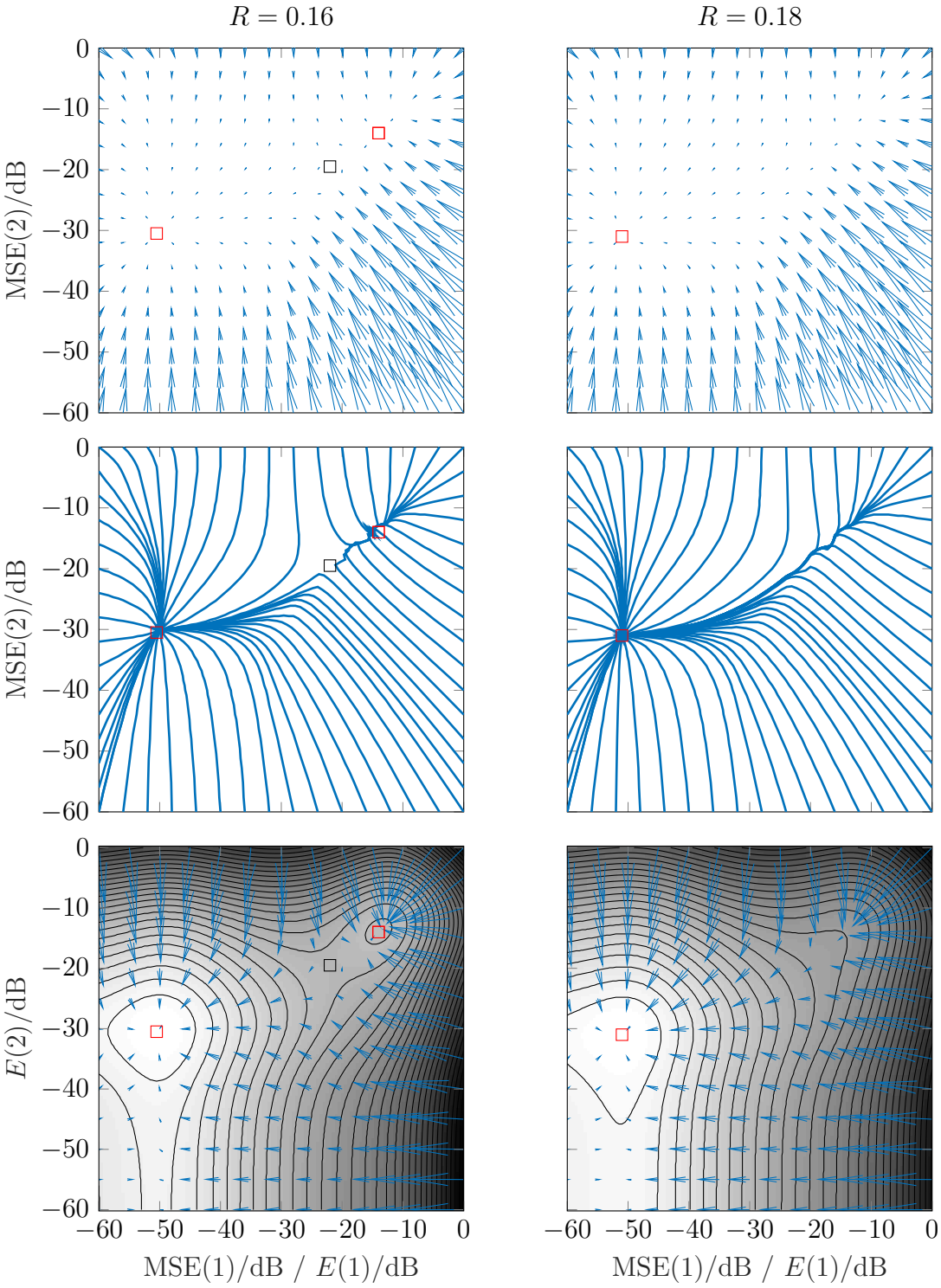


Fig. 3.11 The 2-dimensional (asymmetric) SE and replica analysis MSE prediction ($\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_2$, $\sigma_{\mathsf{w}}^2(1) = -45$dB, $\sigma_{\mathsf{w}}^2(2) = -25$dB). Top row: the SE prediction on the 2-D MSE plane. Middle row: streamline curves whose tangents are the arrows of the top row. Bottom row: 2-D symmetric free energy function with isolines and gradient vectors: the brightness of the shade represents the free energy value (i.e., brighter means larger value). Red/black squares are placed at stable/unstable fixed points and local maxima/saddle points, respectively.

Fully correlated signals

As discussed in Section 3.8, thanks to the decorrelation transform, fully correlated signals ($(\Sigma_{\vec{\mathbf{x}}})_{b,b'} = (\Sigma_{\vec{\mathbf{x}}})_{b,b}$, $b, b' \in 1, 2$) can be interpreted as if there were no correlation ($(\Sigma_{\vec{\mathbf{x}}})_{1,2} = 0$) but zero SNR on one of the measurements, i.e., $(\Sigma_{\vec{\mathbf{w}}})_{2,2} = \infty$. In order to confirm this result by the replica analysis, we plug in $\sigma_{\mathsf{w}}^2(2) \to \infty$ into (3.28): as $\sigma_{\mathsf{w}}^2(2) \to \infty$, $\gamma(2) \to 0$, and the two integrals simplify from a 2-D to a 1-D integral, which is reduced to the 1-D free energy function. Furthermore,

$$
\begin{aligned}
&-\frac{R}{2}\sum_{b=1}^{B}\left(\log\left(2\pi(R\sigma_{\mathsf{w}}^2(b)+E(b))\right)+\frac{\epsilon+R\sigma_{\mathsf{w}}^2(b)}{E(b)+R\sigma_{\mathsf{w}}^2(b)}-\frac{\gamma(b)(1+\epsilon\gamma(b))}{R(1+\gamma(b))}\right)\\
&=-\frac{R}{2}\left(\log\left(2\pi(R\sigma_{\mathsf{w}}^2(1)+E(1))\right)+\frac{\epsilon+R\sigma_{\mathsf{w}}^2(1)}{E(1)+R\sigma_{\mathsf{w}}^2(1)}-\frac{\gamma(1)(1+\epsilon\gamma(1))}{R(1+\gamma(1))}\right)\\
&\quad-\frac{R}{2}\left(\log\left(2\pi(R\sigma_{\mathsf{w}}^2(2)+E(2))\right)+\frac{\epsilon+R\sigma_{\mathsf{w}}^2(2)}{E(2)+R\sigma_{\mathsf{w}}^2(2)}-\frac{\gamma(2)(1+\epsilon\gamma(2))}{R(1+\gamma(2))}\right),
\end{aligned}
$$

where the second term goes to ∞ independently of $E(2)$. This corresponds to a shift in the free energy function $\mathcal{F}(E(1), E(2))$ uniformly towards $-\infty$ along $E(2)$. Since one is interested only in the local extrema (and the curvature), the free energy function is essentially reduced to $\mathcal{F}(E(1))$ (while in fact it is constant over $E(2)$).

Large B limit

Next, the behavior of V-BAMP is discussed as the number of jointly sparse vectors B is increased. In Figure 3.12, the SE curves for different values of B are plotted for the BG prior at nonzero probability $\epsilon = 0.1$, sampling rate $R = 0.25$, and uniform signal and noise $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_B$, $(\Sigma_{\vec{\mathbf{w}}})_{b,b} = -35\text{dB}$, $b \in [B]$. Because of the symmetry, the evolution is characterized by a single effective noise variance $(\Sigma_{\vec{\mathbf{v}}}^{(t)})_{b,b}$. At this rate, all SE curves are crossing the baseline at a relatively low effective noise variance. However, as the rate is further decreased, the first crossing (at a relatively high effective noise variance) that explains the PT appears later for larger values of B. Ultimately, the question arises whether a *first crossing* at a relatively high effective noise variance exists, since the SE curves become nearly parallel to the baseline in a large range of the effective noise variance. This is further illustrated in Figure 3.13: free energy curves for the same setting are plotted at sampling rates around the PT, showing the birth and death of the local extrema as a function of the rate R. One can observe that even though a PT is still observable for $B = 10$, the jump from a relatively high MSE to a relatively low MSE becomes soft. We hypothesize that the appearance of the local

minimum in the free energy curve (and so the second crossing of the SE curve and the baseline) vanishes as B is further increased. It follows that instead of V-BAMP being characterized by the PT rate, the MSE of V-BAMP decreases smoothly with R.

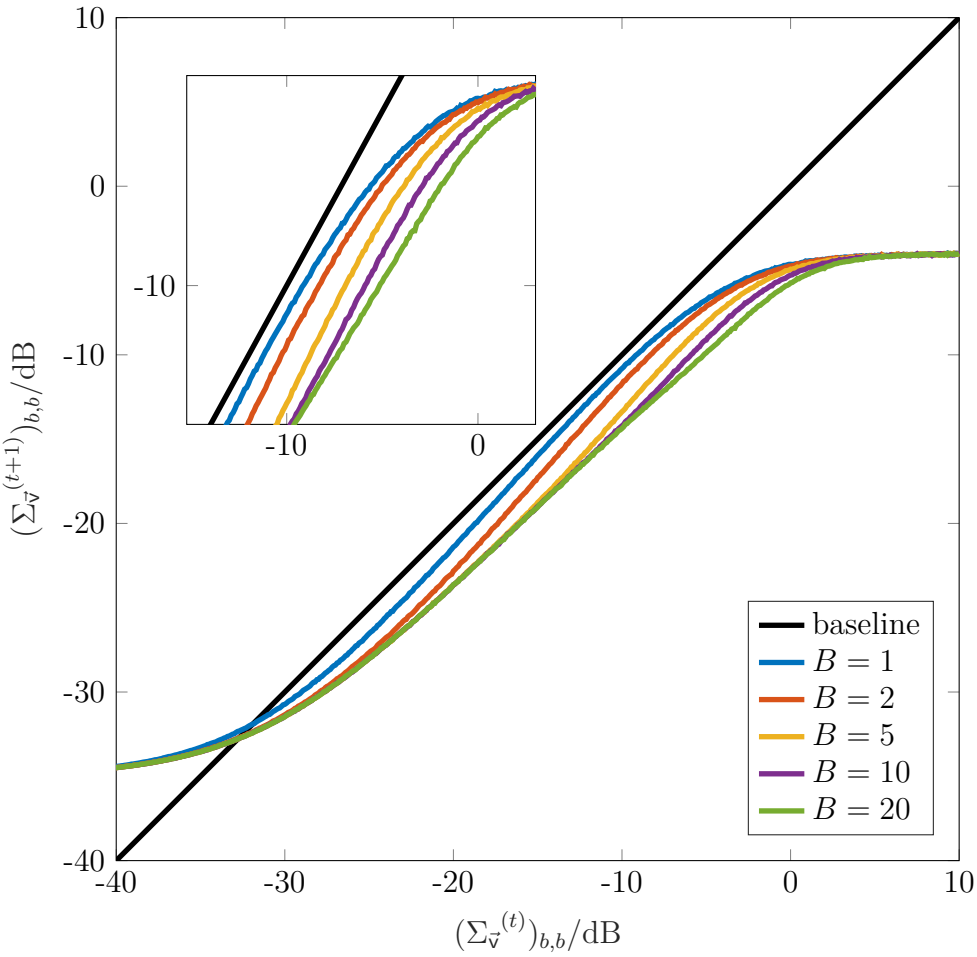


Fig. 3.12 Noisy SE curves for different number of jointly sparse BG signals ($\epsilon = 0.1$, $\Sigma_{\vec{\mathbf{x}}} = \mathbf{I}_B$, $(\Sigma_{\vec{\mathbf{w}}})_{b,b} = -35$dB, $R = 0.25$).

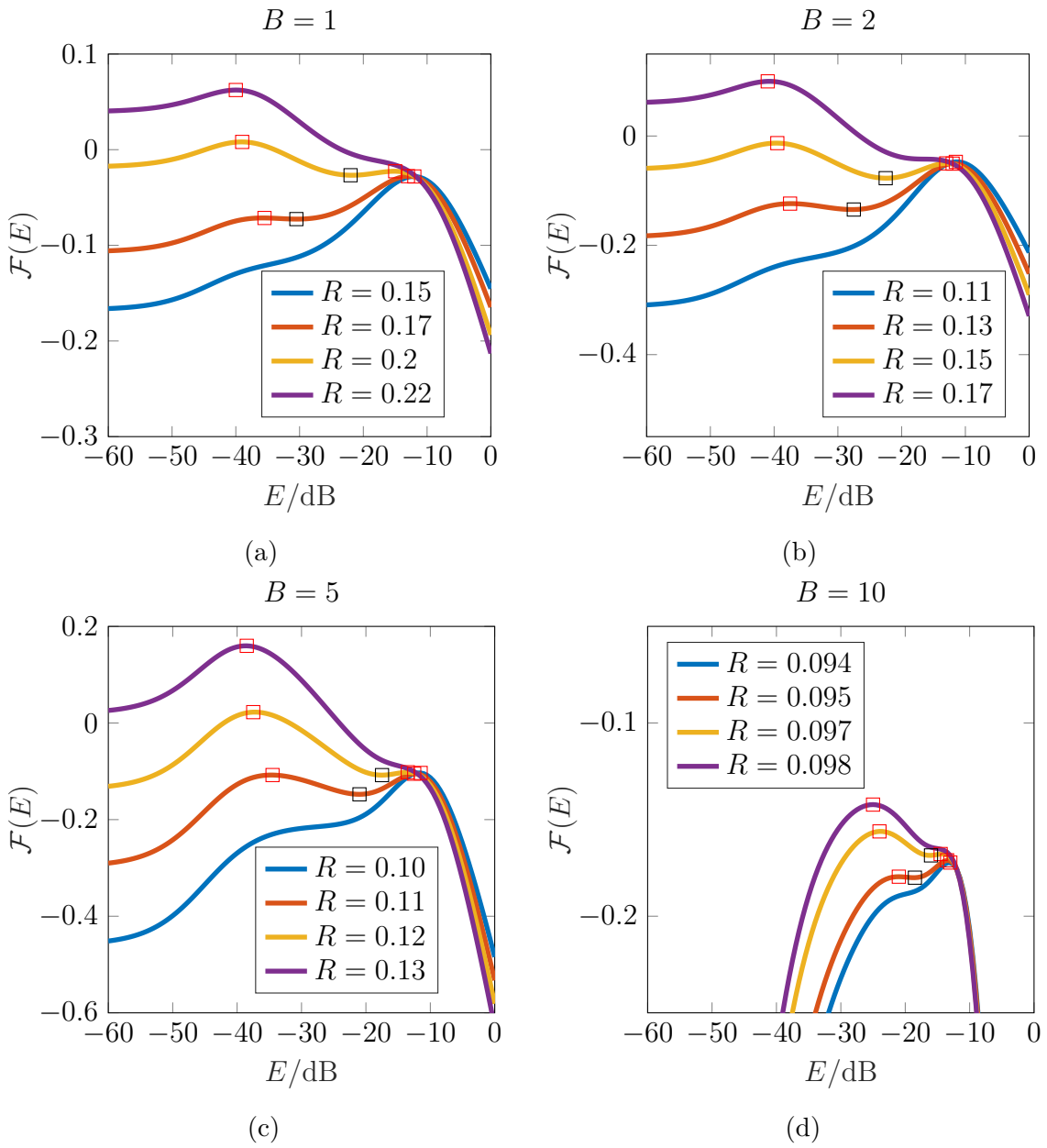


Fig. 3.13 1-D free energy functions for different number of jointly sparse BG signals and different rates ($\epsilon = 0.1$, $(\Sigma_{\vec{\mathbf{w}}})_{b,b} = -35$dB).

Chapter 4

Applications

CS has applications in many fields of signal processing, such as sampling [91], compression [92, 93], solution of inverse problems, design of radiating systems, radar and through-the-wall imaging [7, 94–97], antenna characterization [98], photography [12, 99], and medical imaging [55, 100–104]. The majority of these applications involve sophisticated manipulations that lead to a clean CS formulation (e.g., (1.3) or (1.10)). In the following, we discuss some applications that are already closely related to the linear inverse problem and require only simple modifications in order to uncover the underlying CS problem, and in particular the potential utilization of BAMP and V-BAMP.

4.1 Complex-valued Compressed Sensing

Most of existing work in the field of CS focuses on real-valued signals. However, in many branches of modern signal processing the calculation with complex-valued variables is nearly unavoidable [7, 95, 101, 105]. For an overview of existing algorithms for complex-valued CS the interested reader is referred to [106] and the references therein. Consider a real-valued measurement matrix $\mathbf{A}$ and complex-valued signal (and additive noise) vector. As in the real-valued CS scenario, most components of the unknown $\mathbf{x}$ are 0, and the other components are nonzero in both their real and imaginary parts. That is, the real and imaginary parts of component n are dependent:

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{w}\,, \\ \mathrm{Re}\{\mathbf{y}\} + j\,\mathrm{Im}\{\mathbf{y}\} &= \mathbf{A}\,\mathrm{Re}\{\mathbf{x}\} + j\mathbf{A}\,\mathrm{Im}\{\mathbf{x}\} + \mathrm{Re}\{\mathbf{w}\} + j\,\mathrm{Im}\{\mathbf{w}\}\,. \end{aligned}$$

Assuming the probabilistic measurement model and separating the real and the imaginary parts results in

$$\Big(\mathbf{y}(1), \mathbf{y}(2)\Big) = \mathbf{A}\Big(\mathbf{x}(1), \mathbf{x}(2)\Big) + \Big(\mathbf{w}(1), \mathbf{w}(2)\Big)\,, \tag{4.1}$$

where $\mathbf{y}(1) = \mathrm{Re}\{\mathbf{y}\}$, $\mathbf{x}(1) = \mathrm{Re}\{\mathbf{x}\}$, $\mathbf{w}(1) = \mathrm{Re}\{\mathbf{w}\}$, $\mathbf{y}(2) = \mathrm{Im}\{\mathbf{y}\}$, $\mathbf{w}(2) = \mathrm{Im}\{\mathbf{w}\}$, and $\mathbf{x}(2) = \mathrm{Im}\{\mathbf{x}\}$. When the pdf of the complex-valued signal is expressed as $f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = f_{\mathbf{x}(1),\mathbf{x}(2)}(x_n(1), x_n(2))$ (and the noise is complex normal distributed), together with (4.1) it comprises an MMV scenario, which can be solved efficiently with V-BAMP.

Now assume that not only the signal vector and the measurement noise are complex-valued, but also the measurement matrix $\mathbf{A}$. In general, the real and imaginary parts of the measurement matrix $\mathbf{A}$ are independent, i.e., $\mathrm{Re}\{\mathbf{A}\} = \mathbf{A}(1)$ and $\mathrm{Im}\{\mathbf{A}\} = \mathbf{A}(2)$ are two independent measurement matrices. Separating real and imaginary parts results in

$$\begin{aligned}\mathrm{Re}\{\mathbf{y}\} &= \mathrm{Re}\{\mathbf{A}\}\,\mathrm{Re}\{\mathbf{x}\} - \mathrm{Im}\{\mathbf{A}\}\,\mathrm{Im}\{\mathbf{x}\} + \mathrm{Re}\{\mathbf{w}\}\,,\\ \mathrm{Im}\{\mathbf{y}\} &= \mathrm{Re}\{\mathbf{A}\}\,\mathrm{Im}\{\mathbf{x}\} + \mathrm{Im}\{\mathbf{A}\}\,\mathrm{Re}\{\mathbf{x}\} + \mathrm{Im}\{\mathbf{w}\}\,.\end{aligned}$$

After rewriting (using the notation of (4.1) and $\mathbf{A}(1) = \mathrm{Re}\{\mathbf{A}\}$, $\mathbf{A}(2) = \mathrm{Im}\{\mathbf{A}\}$),

$$\begin{pmatrix}\mathbf{y}(1)\\ \mathbf{y}(2)\end{pmatrix} = \begin{pmatrix}\mathbf{A}(1) & -\mathbf{A}(2)\\ \mathbf{A}(2) & \mathbf{A}(1)\end{pmatrix}\begin{pmatrix}\mathbf{x}(1)\\ \mathbf{x}(2)\end{pmatrix} + \begin{pmatrix}\mathbf{w}(1)\\ \mathbf{w}(2)\end{pmatrix}\,,$$

where the unknown obeys the group sparsity property discussed in Section 1.3.1. While the V-BAMP algorithm presented in Section 3.3 is not directly suited for the group sparse problem, several closely related algorithms have been proposed [47, 48, 107] which cope efficiently with group sparsity.

4.2 Radio Frequency Identification

RFID is a modern technology that allows us to wirelessly identify transponders (tags) with a reader device [108, 109]. For its numerous applications in healthcare, retail, supply chain management, public transport, and many other areas, the interested reader is referred to [60] and references therein. In particular, tags are typically very small, low-cost, and battery-less devices that carry an integrated passive circuit. The reader, when in the vicinity of the tags, initiates the data exchange between the tags

and itself by establishing a wireless link, i.e., emitting a carrier signal. The tags that receive the carrier signal first establish a handshake mechanism with the reader in the *acquisition* phase and then transmit their payload in the *data read-out* phase. The payload typically carries information about the object to which the tag is attached, e.g., product code or sensory information.
Our focus is on the the acquisition phase, which can be modelled by the following sequence of events:

1. The reader emits the known carrier signal.
2. The carrier signal passes through the *forward channel* and is received by the tag.
3. The tag is activated and modulates its signature sequence onto the received signal (*backscatter modulation*) and transmits it [110].
4. The backscatter signal passes through the *backward channel* and is then received by the reader.

In order to formally state the acquisition phase in the baseband signal model, the following quantities are defined:

- N: the total number of tags in the pool. The reader is prepared to acquire and read-out a number of tags from this pool.
- h_n^f: the forward channel coefficient of tag n, i.e., the channel coefficient from the transmit antenna of the reader to tag n. It is typically modelled as a Rayleigh distributed random variable. One can collect the forward channel coefficients into the vector $\mathbf{h}^f = (h_1^f, \ldots, h_N^f)^T$.
- $\mathbf{s}_n$: the signature sequence of tag n. It is typically a sequence of M bits or symbols from a finite symbol alphabet.
- h_n^b: the backward channel coefficient of tag n, i.e., the channel coefficient from tag n to the receive antenna of the reader. It is typically modelled as a Rayleigh distributed random variable. One can collect the backward channel coefficients into the vector $\mathbf{h}^b = (h_1^b, \ldots, h_N^b)^T$.
- $h_n = h_n^f h_n^b$: the compound channel coefficient of tag n. Again, $\mathbf{h} = (h_1, \ldots, h_N)^T$ is the collection of compound channel coefficients.

Suppose K out of N tags from the collection are in the reading range, with indices $n_1, \ldots, n_K$, and respond simultaneously. The forward and backward channel coefficients for the tags not in reading range ($[N] \setminus \{n_1, \ldots, n_K\}$) are assumed to be 0 in the transmission model. Then, the received baseband signal at the reader during acquisition reads

$$\begin{aligned}\mathbf{y} &= \sum_{k=1}^{K} \mathbf{s}_{n_k} h^f_{n_k} h^b_{n_k} + \mathbf{w} \\ &= \mathbf{S}\mathbf{h} + \mathbf{w}\,,\end{aligned}$$

where

$$\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_N)$$

is the collection of the N signature sequences and $\mathbf{w}$ is additive noise, assumed to be white Gaussian. Here, $\mathbf{y}$, $\mathbf{S}$ are known to the reader, w_m ($m \in [M]$) is assumed to be zero-mean i.i.d. Gaussian, and the nonzero components of $\mathbf{h}$ correspond to the indices of the tags in read range. The task of the reader is to determine which of the N tags are in reading range as efficiently and as fast as possible. This could be, e.g., in a shopping cart at checkout or a package of goods at a transportation checkpoint. With clever choice of the signature sequences (e.g., normalized pseudo-random antipodal sequences of sufficient length [21]), the matrix $\mathbf{S}$ fulfills the prerequisits of a valid CS measurement matrix whp. When the collection of all tags is sufficiently large and the number of tags that are to be read out simultaneously is relatively low, i.e., $K \ll N$, the signature sequences can be chosen with length $M \ll N$. Moreover, the channel coefficients can be modelled as random variables, where h^f_n (h^b_n) and $h^f_{n'}$ ($h^b_{n'}$) are i.i.d. and their distribution is known. Taking these factors into consideration, the potential of BAMP is self-evident. The interested reader is referred to [60], in which the application of CS to the presented model is investigated in depth and compared to state-of-the-art methods.

Above, a reader unit with a single receive antenna was considered. Suppose the same setup but with a reader that is equipped with $B > 1$ receive antennas (but still a single carrier emitting transmit antenna). Then, a backward channel coefficient is associated with each pair of tag and receive antenna, i.e., between tag n and receive antenna b one has $h^b_n(b)$. The compound channel coefficient $h^f_n h^b_b(b) = h_n(b)$ is obtained through the chain: transmit antenna $\rightarrow$ tag n $\rightarrow$ receive antenna b. The B received

baseband signals at the reader can be written as

$$\begin{aligned}\Big(\mathbf{y}(1),\ldots,\mathbf{y}(B)\Big) &= \mathbf{S}\,\Big(\mathbf{h}(1),\ldots,\mathbf{h}(B)\Big) + \Big(\mathbf{w}(1),\ldots,\mathbf{w}(B)\Big)\\ &= \mathbf{SH} + \mathbf{W}\,.\end{aligned}$$

Here, each row of $\mathbf{H}$ is either the zero vector or a realization of a random vector representing the collection of B compound channel coefficients, i.e., nonzero whp. That is, the distribution of $\vec{\mathbf{h}}_n = (h(1),\ldots,h(B))^T$ can be determined (based on the physical properties of the system), and the potential of V-BAMP becomes apparent. This model can be straightforwardly extended to the case of multiple readers with an arbitrary number of receive antennas.

In the majority of wireless telecommunication models zero correlation between signals at different receive antennas is approximately guaranteed by the distance between the pairs of antennas, and the resulting signal model describing the system is relatively simple. However, in RFID, due to physical limitations, uncorrelatedness of the signals at the neighboring receiver antennas is not necessarily given. The V-BAMP algorithm, the joint decorrelation transform, and the replica analysis provide an efficient means to cope with the complexity of the signal models that result from closely spaced antennas.

4.3 Multiuser Detection

Wireless multiuser communication systems [111], and specifically machine-to-machine (M2M) communications, have received much attention recently since the number of autonomously communicating devices is expected to grow tremendously [112]. Many practical scenarios can be represented by a star topology, in which a multitude of battery-driven and/or low-complexity devices communicate with one central aggregation node. In the uplink transmission, when the central node collects data from the multitude of spatially distributed devices, several multiple access schemes are available (time/code/frequency division multiple access) for simultaneous transmission such that the receiver is still able to separate the data streams. However, with the growing number of devices in one system, and due to the fact that typically only a small fraction of all devices transmit at the same time, orthogonal multiple access schemes introduce a significant overhead. Applications involving low-complexity and possibly battery-driven devices call for a low-overhead, simple transmission protocol and efficient signal processing methods with minimal cooperation between the users. The interested reader is referred to the related state of the art works [70, 113–116].

4.3.1 Joint Activity Detection and Channel Estimation

Consider a *code division multiple access* (CDMA) wireless communication system with star topology, i.e., N devices (denoted by $1, \ldots, N$) communicate with a central aggregation node C. Addressed is the uplink scenario, in which a subset of the N devices transmit simultaneously to C. The task of activity detection and channel estimation refers to the problem of detecting the set of active (transmitting) users, and estimating their individual channel coefficients. Estimating the channel coefficients is crucial for subsequent data transfer since equalization is an essential part in the transmission chain. For the time being, asynchronicity in the system is neglected and time is discretized into frames: each device is active and transmits within the duration of a complete frame, or is inactive and does not transmit in a given frame. For simplicity, only one frame is considered in the following, because frames do not overlap in time and thus are independent. We assume a flat fading channel, i.e., the channel does not change within the duration of one frame. Within a frame an arbitrary constellation of pilot and data symbols is possible as long as there is at least one pilot symbol per frame. Without loss of generality, one pilot symbol p_n per frame is assumed, which allows for simple presentation. Assume binary phase-shift keying modulation with symbol alphabet $\mathcal{B} = \{-1, 1\}$. The symbols of user n are spread using a unique spreading sequence for user n, $\mathsf{v}_n \in \{-1/\sqrt{M}, 1/\sqrt{M}\}^M$, where $M < N$ and thus the CDMA system is overloaded in the sense that there are more users in the system than signal space dimensions. The transmitted pilot signal corresponding to one symbol of user n is

$$\mathbf{x}_n = p_n \mathbf{s}_n \,.$$

Assuming sporadic device activity, during each frame only a small subset $\mathcal{S} \subset [N]$ of the devices is active, each independently and uniformly with probability $\epsilon \ll 1$. The unknown channel coefficients corresponding to the active devices are assumed to be complex, i.e., $h_n = \mathrm{Re}\{h_n\} + j\,\mathrm{Im}\{h_n\} \in \mathbb{C}$ for $n \in \mathcal{S}$, whereas the channel coefficients corresponding to the inactive devices are defined to be zero, i.e., $h_n = 0$ for $n \notin \mathcal{S}$. The pilot signal received by the central aggregation node C from device n becomes

$$\mathbf{y}_n = h_n \mathbf{x}_n \,, \quad n \in [N] \,,$$

which sum up to the overall received signal

$$\mathbf{y} = \sum_{n \in \mathcal{S}} \mathbf{y}_n + \mathbf{w} \,, \tag{4.2}$$

with $\mathsf{w}_m \sim \mathcal{CN}(0, 2\sigma_{\mathsf{w}}^2)$ ($m \in [M]$) being complex zero-mean *additive white Gaussian noise* (AWGN) i.i.d. over m. Rewrite (4.2) as

$$\begin{aligned} \mathbf{y} &= \mathbf{S}\,\mathrm{diag}(p_1, \ldots, p_N)\mathbf{h} + \mathbf{w} \\ &= \tilde{\mathbf{S}}\mathbf{h} + \mathbf{w}\,, \end{aligned} \tag{4.3}$$

where $\tilde{\mathbf{S}} = (p_1\mathbf{s}_1, \ldots, p_N\mathbf{s}_N)$. The task of simultaneous activity detection and channel estimation amounts to detecting and estimating the nonzero channel coefficients. In (4.3), the unknown $\mathbf{h}$ is complex-valued and sparse, and with proper choice of pilot symbols and CDMA sequences the matrix $\tilde{\mathbf{S}}$ is a valid CS measurement matrix fulfilling the conditions for robust sampling (cf. Section 1.1.4). As shown in Section 4.1, (4.3) is an MMV CS measurement, which, when the pdf of the channel coefficients is available, can be solved via V-BAMP. Remember that V-BAMP is an approximate MMSE estimator and the estimate $\hat{h}_n$ almost never equals exactly h_n. Activity detection consists in detecting the indices of the nonzeros in $\mathbf{h}$. When the number of active users is known, e.g., K, this can be achieved by keeping the K indices with largest magnitude h_n. However, when the number of active users is not known, other techniques such as thresholding or the EM algorithm (see Section 3.5) can be employed. The similarity of the above setup to the RFID application is clear. However, in wireless multiuser communication systems activity detection and channel estimation can be combined with (subsequent) data transmission, as discussed in the following. Pilot and data symbols can be arranged in particular patterns into frames. Performing simultaneous frame-wise activity detection, channel estimation, and data demodulation is a potential application of V-BAMP, with a prior pdf that is matched to the frame composed of pilot and data symbols.

4.3.2 QAM Demodulation

In the previous section channel estimation using pilot symbols was considered. Once the channel coefficients are available at the central receiving node, it is ready to receive the user data payload as it can equalize the channel within the channel coherence time [117].

Only one frame is considered in the following. User n is either inactive and not transmitting, or is active and transmitting $\log_2 U$ bits in a frame, that are Gray-mapped onto an Q-ary QAM symbol. In QAM, which is well established in practice, the QAM symbols are composed of a real and an imaginary part. Both take values

from the same real-valued alphabet, which is symmetric around 0:

$$s^{(q)*} = a + jb\,, \quad a, b \in \mathcal{A} = \frac{1}{Z}\{\pm 1, \pm 3, \ldots, \pm(\sqrt{Q} - 1)\}$$

with Z being a normalization factor depending on the modulation order and $\mathcal{A}$ being the set of real values for the real/imaginary parts of the complex-valued symbol alphabet $\mathcal{S}^*$. In practice, $U \in \{4, 16, 64, 256, 1024, 4096\}$. Each of the QAM symbols $s^{(q)*} \in \mathcal{S}^*$ is represented by a complex number, i.e., $s^{(q)*} = \mathrm{Re}\{s^{(q)*}\} + j\,\mathrm{Im}\{s^{(q)*}\}$, whose real and imaginary parts correspond to the in-phase and quadrature-phase components of the symbols. Note that the superscript $*$ denotes complex-valued quantities. The sporadic node activity is modeled by the activity probability ϵ, i.e., in a given frame each node transmits with probability ϵ independently from the other nodes. When the inactive users are represented as transmitting a zero symbol, the vector of transmitted symbols $\mathbf{x}^* = (x_1^*, \ldots, x_N^*)^T$ consists of elements from the extended alphabet

$$x_n^* \in \mathcal{S}^* \cup \{0\} \quad \forall n \in [N]\,,$$

where $\mathcal{S}^* = \{s^{(1)*}, \ldots, s^{(Q)*}\}$ denotes the QAM symbol alphabet normalized to average symbol energy 1. To keep the theoretical description simple, only one transmit symbol per frame is assumed. The canonical input-output relationship of the uplink transmission from the N nodes to the center node C can be cast as

$$\mathbf{y}^* = \mathbf{A}\mathbf{x}^* + \mathbf{w}^*\,, \tag{4.4}$$

where $\tilde{\mathbf{w}} \in \mathbb{C}^K$ is circularly-symmetric complex AWGN, and the *measurement matrix* $\mathbf{A} \in \mathbb{R}^{M \times N}$ subsumes the transmit and receive filters, the channel impulse responses, and the multiple access scheme signature sequences [70]. In orthogonal systems, (4.4) is well determined. However, in very large communication systems one can fall back to the underdetermined design, i.e., $M < N$, which can be handled by the CS framework. One can write (4.4) equivalently as

$$\Big(\mathbf{y}(1),\, \mathbf{y}(2)\Big) = \mathbf{A}\Big(\mathbf{x}(1),\, \mathbf{x}(2)\Big) + \Big(\mathbf{w}(1),\, \mathbf{w}(2)\Big)\,, \tag{4.5}$$

where $\vec{\mathbf{x}}_n = (x_n(1),\, x_n(2))^T$, and the equivalent QAM vector alphabet is

$$\begin{aligned}\mathcal{S} &= \Big\{\Big(\mathrm{Re}\{s^{(1)*}\}, \mathrm{Im}\{s^{(1)*}\}\Big)^T, \ldots, \Big(\mathrm{Re}\{s^{(Q)*}\}, \mathrm{Im}\{s^{(Q)*}\}\Big)\Big\} \cup \{(0,0)^T\} \\ &= \{\vec{\mathbf{s}}^{(1)}, \ldots, \vec{\mathbf{s}}^{(Q)}\}\,.\end{aligned}$$

If one assumes that the distribution of the transmitted symbols is uniform, the pdf of $\vec{\mathbf{x}}_n$ becomes

$$f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}_n) = (1-\epsilon)\delta(\vec{\mathbf{x}}_n) + \frac{\epsilon}{Q}\sum_{q=1}^{Q}\delta(\vec{\mathbf{x}}_n - \vec{\mathbf{s}}^{(q)})\,.$$

If the matrix $\mathbf{A}$ is suitably designed (which is partially ensured by the randomness of the channel coefficients), together with (4.5) this constitutes a probabilistic MMV scenario, which can be solved efficiently with V-BAMP.

Chapter 5

Conclusions

This thesis investigated the V-BAMP algorithm for the DCS and MMV scenarios, which can account for arbitrary signal and noise correlation. The effects of correlation in the signal were examined both using numerical simulations and the SE equations. A joint decorrelation method showed that an arbitrary MMV measurement can be transformed (using an invertible linear transformation) into an equivalent measurement without signal and noise correlations, such that the noise covariance matrix contains the B individual inverse SNRs on its diagonal. Additionally, it was proven that the V-BAMP algorithm and its SE are equivariant w.r.t. invertible linear transformations. That is, both V-BAMP and its theoretical analysis on the original measurement and the transformed (decorrelated) measurement are equivalent. Furthermore, we show that for the widely employed BG prior V-BAMP preserves the diagonality of the effective noise covariance matrix. An important consequence is that while in general V-BAMP is described by $B(B+1)/2$ states for B jointly measured vectors, for BG signals this is reduced to only B states, which are in direct correspondence with the B individual MSEs. Furthermore, when performing the analysis of V-BAMP with BG signals, it suffices to account for the set of measurements with uncorrelated signal and noise. This allows us to employ the replica trick borrowed from statistical physics to derive the MMSE for the CS measurement. This work shows that the underlying V-BAMP dynamics predicted by the replica method matches those predicted by SE, and supports the hypothesis that V-BAMP can be interpreted as a gradient ascent on the B-dimensional free energy function of the MSEs. Together with the joint decorrelation transform, this provides an in-depth analysis of V-BAMP in terms of the MSE evolution for the multivariate BG signal prior in the jointly sparse CS scenario. Moreover, it was shown that as the number of measured jointly sparse vectors increases,

the phase transition behavior of V-BAMP vanishes, and that instead the estimation error decreases smoothly with increasing sampling rate.

Appendix A

MMSE Estimator: Derivative and (Co-)Variance Relation

Scalar Case

Given a realization x of a random variable x with pdf $f_{\mathsf{x}}(x)$ and its noisy observation

$$u = x + w$$

with $\mathsf{w} \sim \mathcal{N}(0, \sigma_{\mathsf{w}}^2)$ being independent additive Gaussian noise, its MMSE estimator is

$$\hat{x}(u, \sigma_{\mathsf{w}}^2) = \mathrm{E}\left\{\mathsf{x} \mid \mathsf{u} = u, \sigma_{\mathsf{w}}^2\right\} .$$

Then, the following relation holds:

$$\mathrm{Var}\left\{\mathsf{x} \mid \mathsf{u} = u, \sigma_{\mathsf{w}}^2\right\} = \sigma_{\mathsf{w}}^2 \frac{d}{du} \hat{x}(u, \sigma_{\mathsf{w}}^2) .$$

The proof is provided in the next section for a more general form of this statement.

Multivariate Case

Given a realization $\mathbf{x}$ of a random vector $\boldsymbol{\mathsf{x}} \in \mathbb{R}^N$ with pdf $f_{\boldsymbol{\mathsf{x}}}(\mathbf{x})$ and its noisy observation

$$\mathbf{u} = \mathbf{x} + \mathbf{w}$$

with $\boldsymbol{\mathsf{w}} \sim \mathcal{N}(0, \Sigma_{\vec{\boldsymbol{\mathsf{w}}}})$ being independent additive Gaussian noise, its MMSE estimator is

$$\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\boldsymbol{\mathsf{w}}}}) = \mathrm{E}\left\{\boldsymbol{\mathsf{x}} \mid \boldsymbol{\mathsf{u}} = \mathbf{u}, \Sigma_{\vec{\boldsymbol{\mathsf{w}}}}\right\} .$$

Then, the following relation holds:

$$\mathrm{Cov}\left\{\mathbf{x} \mid \mathbf{u} = \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\} = \frac{d}{d\mathbf{u}^T}\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}})\Sigma_{\vec{\mathbf{w}}}\,.$$

Proof: Given the definition of the conditional mean and covariance,

$$\begin{aligned}
\mathrm{E}\left\{\mathbf{x} \mid \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\} &= \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x} f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\
\mathrm{Cov}\left\{\mathbf{x} \mid \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\} &= \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x}\mathbf{x}^T f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} - \mathrm{E}\left\{\mathbf{x} \mid \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\}\mathrm{E}\left\{\mathbf{x} \mid \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\}^T\,,
\end{aligned}$$

we have

$$\begin{aligned}
\frac{d}{d\mathbf{u}^T}\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}})\Sigma_{\vec{\mathbf{w}}} =& \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x} f_{\mathbf{x}}(\mathbf{x}) \frac{d}{d\mathbf{u}^T} f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) d\mathbf{x}\Sigma_{\vec{\mathbf{w}}} \\
&- \int_{\mathbb{R}^N} \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\mathbf{x} f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}\ \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\frac{d}{d\mathbf{u}^T} f_{\mathbf{u}}(\mathbf{u})\Sigma_{\vec{\mathbf{w}}}\,. \qquad \text{(A.1)}
\end{aligned}$$

Since $f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) = \mathcal{N}(\mathbf{u}; \mathbf{x}, \Sigma_{\vec{\mathbf{w}}})$ [118],

$$\frac{d}{d\mathbf{u}^T} f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) = f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x})(\mathbf{x} - \mathbf{u})^T \Sigma_{\vec{\mathbf{w}}}^{-1}\,. \qquad \text{(A.2)}$$

Furthermore, the MMSE estimator can also be written as [119, 120]

$$\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}}) = \mathbf{u} + \Sigma_{\vec{\mathbf{w}}}\frac{1}{f_{\mathbf{u}}(\mathbf{u})}\frac{d}{d\mathbf{u}} f_{\mathbf{u}}(\mathbf{u})\,. \qquad \text{(A.3)}$$

Combining (A.1), (A.2), and (A.3) we have

$$\begin{aligned}
\frac{d}{d\mathbf{u}^T}\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}})\Sigma_{\vec{\mathbf{w}}} &= \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x}(\mathbf{x} - \mathbf{u})^T f_{\mathbf{u}|\mathbf{x}}(\mathbf{x} \mid \mathbf{u}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\
&\quad - \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x} f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}(\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}}) - \mathbf{u})^T \\
&= \frac{1}{f_{\mathbf{u}}(\mathbf{u})}\int_{\mathbb{R}^N} \mathbf{x}\mathbf{x}^T f_{\mathbf{u}|\mathbf{x}}(\mathbf{u} \mid \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} - \hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}})\hat{\mathbf{x}}(\mathbf{u}, \Sigma_{\vec{\mathbf{w}}})^T \\
&= \mathrm{Cov}\left\{\mathbf{x} \mid \mathbf{u}, \Sigma_{\vec{\mathbf{w}}}\right\}\,,
\end{aligned}$$

which completes the proof.

Appendix B

Equivariance of MMV VBAMP and its SE

Consider Algorithm 2 with the transformed variables $\Sigma_{\tilde{\vec{\mathbf{x}}}}$, $\mathbf{T}\hat{\vec{\mathbf{x}}}_n^{(t)}$, $\mathbf{T}\vec{\mathbf{r}}_m^{(t)}$, $\mathbf{T}\vec{\mathbf{u}}_n^{(t)}$, $\Sigma_{\tilde{\vec{\mathbf{v}}}}^{(t)}$. Lines 5 and 6 are trivially equivariant. The equivariance of line 7 can be verified by the invariance property of MMSE estimators to affine transformations [63, Ch. 11.4]. In the residual term (line 8), the equivariance of $\vec{\mathbf{y}}_m - (\mathbf{A}(1)\hat{\mathbf{x}}^{(t)}(1), \ldots, \mathbf{A}(B)\hat{\mathbf{x}}^{(t)}(B))_m$ is trivial. It remains to show that the Onsager term is equivariant. Thus, the transformed Onsager term is written as

$$
\begin{aligned}
&\frac{1}{M}\sum_{n=1}^{N} F'(\mathbf{T}\vec{\mathbf{u}}_n; \mathbf{T}\Sigma_{\vec{\mathbf{v}}}\mathbf{T}^T)\mathbf{T}\vec{\mathbf{r}}_m \\
&\overset{(1)}{=} \frac{1}{M}\sum_{n=1}^{N} \mathrm{Cov}\{\tilde{\vec{\mathbf{x}}} \mid \mathbf{T}\vec{\mathbf{u}}_n; \mathbf{T}\Sigma_{\vec{\mathbf{v}}}\mathbf{T}^T\}\left(\mathbf{T}\Sigma_{\vec{\mathbf{v}}}\mathbf{T}^T\right)^{-1}\mathbf{T}\vec{\mathbf{r}}_m \\
&= \frac{1}{M}\sum_{n=1}^{N} \mathrm{E}\{\langle \tilde{\vec{\mathbf{x}}} - \mathrm{E}\{\tilde{\vec{\mathbf{x}}}\}\rangle \mid \mathbf{T}\vec{\mathbf{u}}_n; \mathbf{T}\Sigma_{\vec{\mathbf{v}}}\mathbf{T}^T\}\mathbf{T}^{-T}\Sigma_{\vec{\mathbf{v}}}^{-1}\vec{\mathbf{r}}_m \\
&= \frac{1}{M}\sum_{n=1}^{N} \mathbf{T}\,\mathrm{E}\{\langle \vec{\mathbf{x}} - \mathrm{E}\{\vec{\mathbf{x}}\}\rangle \mid \vec{\mathbf{u}}_n; \Sigma_{\vec{\mathbf{v}}}\}\mathbf{T}^T\mathbf{T}^{-T}\Sigma_{\vec{\mathbf{v}}}^{-1}\vec{\mathbf{r}}_m \\
&\overset{(2)}{=} \mathbf{T}\frac{1}{M}\sum_{n=1}^{N} \mathrm{Cov}\{\vec{\mathbf{x}} \mid \vec{\mathbf{u}}_n; \Sigma_{\vec{\mathbf{v}}}\}\Sigma_{\vec{\mathbf{v}}}^{-1}\vec{\mathbf{r}}_m \\
&= \mathbf{T}\frac{1}{M}\sum_{n=1}^{N} F'(\vec{\mathbf{u}}_n; \Sigma_{\vec{\mathbf{v}}})\vec{\mathbf{r}}_m\,,
\end{aligned}
$$

where (1) and (2) follow from the result in Appendix A, and the proof of part 1) of Theorem 2 is complete. Using elementary probability theory and the invariance

property of MMSE estimators to affine transformations [63, Ch. 11.4], part 2) of Theorem 2 can be proven by simple calculation.

Appendix C

Diagonality of SE with BG Prior

We show that MMV SE (3.10) preserves diagonality for the BG prior. In particular, we prove that if $\Sigma_{\vec{\mathbf{v}}}^{(t)}$, $\Sigma_{\vec{\mathbf{w}}}$ and $\Sigma_{\vec{\mathbf{x}}}$ are diagonal, then

$$\Sigma_{\vec{\mathbf{v}}}^{(t+1)} = \Sigma_{\vec{\mathbf{w}}} + \frac{1}{R} \underbrace{\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}} \left\{ \langle F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^t; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}} \rangle \right\}}_{\mathbf{C}}$$

is also diagonal. Since the factor $\frac{1}{R}$ and the term $\Sigma_{\vec{\mathbf{w}}}$ do not influence the diagonality, we examine the expectation $\mathbf{C}$. Writing out the expectation we have

$$\begin{aligned}
\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \langle F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}} \rangle \right\} &= \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \langle F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) \rangle \right\} \\
&\quad - \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \langle \vec{\mathbf{x}} \rangle \right\} \\
&\quad - \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) \vec{\mathbf{x}}^T \right\} \\
&\quad - \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \vec{\mathbf{x}} F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)})^T \right\}
\end{aligned}$$

Using the argument that the mapping $\vec{\mathbf{x}} \to \mathrm{E}_{\vec{\mathbf{x}}}\{\vec{\mathbf{x}}\}$ is self-adjoint and substituting the conditional expectation for the estimator $F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)})$, we have

$$\begin{aligned}
\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) \vec{\mathbf{x}}^T \right\} &= \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \vec{\mathbf{x}} \mid \vec{\mathrm{x}} + \vec{\mathrm{v}}, \Sigma_{\vec{\mathbf{v}}}^{(t)} \right\} \vec{\mathbf{x}}^T \right\} \\
&= \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \vec{\mathbf{x}} \vec{\mathbf{x}}^T \right\} \mid \vec{\mathrm{x}} + \vec{\mathrm{v}}, \Sigma_{\vec{\mathbf{v}}}^{(t)} \right\} \\
&= \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \vec{\mathbf{x}} \mid \vec{\mathrm{x}} + \vec{\mathrm{v}}, \Sigma_{\vec{\mathbf{v}}}^{(t)} \right\} \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \vec{\mathbf{x}}^T \mid \vec{\mathrm{x}} + \vec{\mathrm{v}}, \Sigma_{\vec{\mathbf{v}}}^{(t)} \right\} \right\} \\
&= \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}} \left\{ \langle F(\vec{\mathbf{x}} + \vec{\mathbf{v}}^{(t)}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)}) \rangle \right\} .
\end{aligned}$$

Overall, we have

$$\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}}^{(t)};\Sigma_{\vec{\mathbf{v}}}{}^{(t)})-\vec{\mathbf{x}}\rangle\right\}=\mathrm{E}_{\vec{\mathbf{x}}}\left\{\langle\vec{\mathbf{x}}\rangle\right\}-\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}^{(t)}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}}^{(t)};\Sigma_{\vec{\mathbf{v}}}{}^{(t)})\rangle\right\}.$$

Using the substitution $\vec{\mathbf{u}}^{(t)}=\vec{\mathbf{x}}+\vec{\mathbf{v}}^{(t)}$ and writing out the integral defined by the expectation gives

$$\begin{aligned}\mathbf{C}_{i,j}&=\int_{\mathbb{R}^B}(\vec{\mathbf{x}})_i(\vec{\mathbf{x}})_j f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}})d\vec{\mathbf{x}}-\int_{\mathbb{R}^B}F(\vec{\mathbf{u}}^{(t)};\Sigma_{\vec{\mathbf{v}}}^{(t)})_i F(\vec{\mathbf{u}}^{(t)};\Sigma_{\vec{\mathbf{v}}}^{(t)})_j f_{\vec{\mathbf{u}}^{(t)}}(\vec{\mathbf{u}}^{(t)})d\vec{\mathbf{u}}\\&=(\Sigma_{\vec{\mathbf{x}}})_{i,j}-\int_{\mathbb{R}^B}F(\vec{\mathbf{u}}^{(t)};\Sigma_{\vec{\mathbf{v}}}^{(t)})_i F(\vec{\mathbf{u}}^{(t)};\Sigma_{\vec{\mathbf{v}}}^{(t)})_j f_{\vec{\mathbf{u}}^{(t)}}(\vec{\mathbf{u}}^{(t)})d\vec{\mathbf{u}}\end{aligned}$$

where $f_{\vec{\mathbf{u}}^{(t)}}(\vec{\mathbf{u}}^{(t)})=f_{\vec{\mathbf{x}}}(\vec{\mathbf{u}}^{(t)})*f_{\vec{\mathbf{v}}^{(t)}}(\vec{\mathbf{u}}^{(t)})=(1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}^{(t)};\mathbf{0},\Sigma_{\vec{\mathbf{v}}}^{(t)})+\epsilon\mathcal{N}(\vec{\mathbf{u}}^{(t)};\mathbf{0},\Sigma_{\vec{\mathbf{u}}}^{(t)})$, and with $\Sigma_{\vec{\mathbf{u}}}^{(t)}=\Sigma_{\vec{\mathbf{x}}}+\Sigma_{\vec{\mathbf{v}}}^{(t)}$, which is diagonal by assumption. By assumption, $(\Sigma_{\vec{\mathbf{x}}})_{i,j}=0$ for $i\neq j$. Substituting the estimator function (3.8) into the second term we can verify that the integrand has odd symmetry w.r.t. $(\vec{\mathbf{u}})_i$ and thus integrates to 0 if $i\neq j$. Thus, if $\Sigma_{\vec{\mathbf{w}}}$, $\Sigma_{\vec{\mathbf{x}}}$, and $\Sigma_{\vec{\mathbf{v}}}^{(t)}$ are diagonal, $\Sigma_{\vec{\mathbf{v}}}^{(t+1)}$ is diagonal as well.

Appendix D

SE Integral Evaluation

The SE equation for (MMV) (V)BAMP reads

$$\Sigma_{\vec{\mathbf{v}}}^{(t+1)} = \Sigma_{\vec{\mathbf{w}}} + \frac{1}{R}\,\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\}\,.$$

with pdfs $f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}})$ and $f_{\vec{\mathbf{v}}}(\vec{\mathbf{v}}) = \mathcal{N}(\vec{\mathbf{v}};\mathbf{0},\Sigma_{\vec{\mathbf{v}}}^{(t)})$. Given that $\vec{\mathbf{x}},\vec{\mathbf{v}} \in \mathbb{R}^B$, the expectation

$$\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\}$$

requires in general $B(B+1)$ integrals (using the fact that the covariance matrix is symmetric) over $B+B$ variables. When the integrals are not available in closed form, alternatively, using Monte Carlo simulation, a sufficiently large number of pseudo-random vectors $\vec{\mathbf{x}}^i$ and $\vec{\mathbf{v}}^i$ $(i = 1,\ldots,I)$ can be generated independently according to the pdfs $f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}})$ and $f_{\vec{\mathbf{v}}}(\vec{\mathbf{v}})$. Then, the expectation can be estimated by

$$\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\} \approx \frac{1}{I}\sum_{i=1}^{I}\langle F(\vec{\mathbf{x}}^i+\vec{\mathbf{v}}^i;\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}^i\rangle\,.$$

This, however, can become computationally expensive, as the accuracy of the estimate requires (especially at low noise levels) I to be very large. This is partly due to the fact that both $\vec{\mathbf{x}}$ and $\vec{\mathbf{v}}$ need a very large number of independent realizations. By using the calculation derived in Appendix C, we have

$$\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)}) - \vec{\mathbf{x}}\rangle\right\} = \mathrm{E}_{\vec{\mathbf{x}}}\left\{\langle\vec{\mathbf{x}}\rangle\right\} - \mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}}+\vec{\mathbf{v}};\Sigma_{\vec{\mathbf{v}}}{}^{(t)})\rangle\right\}\,.$$

For the BG prior $f_{\vec{\mathbf{x}}}(\vec{\mathbf{x}}) = (1-\epsilon)\delta(\vec{\mathbf{x}}) + \epsilon\mathcal{N}(\vec{\mathbf{x}}; \mathbf{0}, \Sigma_{\vec{\mathbf{x}}})$,

$$\mathrm{E}_{\vec{\mathbf{x}}}\left\{\langle\vec{\mathbf{x}}\rangle\right\} = \epsilon\Sigma_{\vec{\mathbf{x}}}$$

and with $\vec{\mathbf{u}} = \vec{\mathbf{x}} + \vec{\mathbf{v}}$,

$$\mathrm{E}_{\vec{\mathbf{x}},\vec{\mathbf{v}}}\left\{\langle F(\vec{\mathbf{x}} + \vec{\mathbf{v}}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)})\rangle\right\} = \mathrm{E}_{\vec{\mathbf{u}}}\left\{\langle F(\vec{\mathbf{u}}; \Sigma_{\vec{\mathbf{v}}}{}^{(t)})\rangle\right\} ,$$

where $f_{\vec{\mathbf{u}}}(\vec{\mathbf{u}}) = f_{\vec{\mathbf{x}}}(\vec{\mathbf{u}}) * f_{\vec{\mathbf{v}}}(\vec{\mathbf{u}}) = (1-\epsilon)\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{v}}}^{(t)}) + \epsilon\mathcal{N}(\vec{\mathbf{u}}; \mathbf{0}, \Sigma_{\vec{\mathbf{u}}}^{(t)})$ and $\Sigma_{\vec{\mathbf{u}}}^{(t)} = \Sigma_{\vec{\mathbf{x}}} + \Sigma_{\vec{\mathbf{v}}}^{(t)}$. Thus, it suffices to perform the Monte Carlo integration only over B dimensions variable, namely the components of $\vec{\mathbf{u}}^i$.

List of Acronyms

AMP approximate message passing

AWGN additive white Gaussian noise

BAMP Bayesian approximate message passing

BG Bernoulli-Gauss

BP basis pursuit

CDMA code division multiple access

CS compressed sensing

DCS distributed compressed sensing

EM expectation-maximization

i.i.d. independent and identically distributed

JSM joint sparsity model

LASSO least absolute shrinkage and selection operator

MAP maximum a posteriori

MMSE minimum mean squared error

MMV multiple measurement vectors

MSE mean squared error

pdf probability density function

PT phase transition

PTC phase transition curve

QAM quadrature amplitude modulation

RFID radio-frequency identification

RIP restricted isometry property

SE state evolution

SNR signal-to-noise ratio

V-BAMP vector Bayesian approximate message passing

whp with high probability

References

[1] M. Castells, *The Rise of The Network Society: The Information Age: Economy, Society and Culture*, ser. Information Age Series. Wiley, 2000, no. v. 1.

[2] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W. W. Norton, 2014.

[3] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectrum*, vol. 34, no. 6, pp. 52–59, 1997.

[4] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[5] K. Brandenburg, "MP3 and AAC explained," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding.* Audio Engineering Society, 1999.

[6] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[7] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *IEEE Radar Conference.* IEEE, 2007, pp. 128–133.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[10] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," *Preprint*, 2011.

[11] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[12] M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[13] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing.* Prentice Hall, 2000.

[14] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[15] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer New York, 2010. [Online]. Available: https://books.google.at/books?id=d5b6lJI9BvAC

[16] A. M. Tillmann and M. E. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1248–1259, 2014.

[17] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[18] M. Fornasier and H. Rauhut, "Compressive sensing," in *Handbook of Mathematical Methods in Imaging.* Springer, 2011, pp. 187–228.

[19] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing.* Birkhäuser Basel, 2013, vol. 1, no. 3.

[20] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[21] M. Mayer, N. Görtz, and J. Kaitovic, "RFID tag acquisition via compressed sensing," in *2014 IEEE RFID Technology and Applications Conference (RFID-TA).* IEEE, 2014, pp. 26–31.

[22] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 389–400, 2013.

[23] J. A. Tropp, M. B. Wakin, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Random filters for compressive sampling and reconstruction," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3. IEEE, 2006, pp. III–III.

[24] S. Qaisar, R. M. Bilal, W. Iqbal, M. Naureen, and S. Lee, "Compressive sensing: From theory to applications, a survey," *Journal of Communications and Networks*, vol. 15, no. 5, pp. 443–456, 2013.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[26] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[28] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[29] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[30] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[31] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[32] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 330–341, 2010.

[33] K. K. Herrity, A. C. Gilbert, and J. A. Tropp, "Sparse approximation via iterative thresholding," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3. IEEE, 2006, pp. III–III.

[34] M. Fornasier and H. Rauhut, "Iterative thresholding algorithms," *Applied and Computational Harmonic Analysis*, vol. 25, no. 2, pp. 187–208, 2008.

[35] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[36] A. Montanari, "Graphical models concepts in compressed sensing," *Compressed Sensing: Theory and Applications*, pp. 394–438, 2012.

[37] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, 2001.

[38] A. Maleki, *PhD Thesis: Approximate Message Passing Algorithms for Compressed Sensing*. Stanford University, 2010.

[39] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[40] ——, "Message passing algorithms for compressed sensing: I. motivation and construction," in *IEEE Workshop on Information Theory*. IEEE, 2010, pp. 1–5.

[41] ——, "How to design message passing algorithms for compressed sensing," *preprint*, 2011.

[42] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.

[43] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[44] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[45] R. Chartrand and B. Wohlberg, "A nonconvex ADMM algorithm for group sparsity with sparse groups," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6009–6013.

[46] J. Huang and T. Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[47] P. Schniter, "Turbo reconstruction of structured sparse signals," in *44th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2010, pp. 1–6.

[48] M. Mayer and N. Goertz, "Bayesian optimal approximate message passing to recover structured sparse signals," *arXiv preprint arXiv:1508.01104*, 2015.

[49] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[50] S. Sarvotham, D. Baron, M. Wakin, M. F. Duarte, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Conference Record of The 39th Asilomar Conference on Signals, Systems, and Computers*, 2005, pp. 1537–1541.

[51] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 340–354, 2013.

[52] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[53] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.

[54] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *The Annals of Statistics*, pp. 1–47, 2011.

[55] D. Liang, L. Ying, and F. Liang, "Parallel MRI acceleration using M-FOCUSS," in *3rd International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 2009, pp. 1–4.

[56] O. Lee, J. M. Kim, Y. Bresler, and J. C. Ye, "Compressive diffuse optical tomography: noniterative exact reconstruction using joint sparsity," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1129–1142, 2011.

[57] M. A. Kanso and M. G. R., "Compressed RF tomography for wireless sensor networks: Centralized and decentralized approaches," in *International Conference on Distributed Computing in Sensor Systems*. Springer, 2009, pp. 173–186.

[58] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, 1995.

[59] G. Tzagkarakis, D. Milioris, and P. Tsakalides, "Multiple-measurement Bayesian compressed sensing using GSM priors for DOA estimation," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 2610–2613.

[60] M. Mayer, *PhD Thesis: Radio Frequency Identification with Compressed Sensing*. Technische Universität Wien, 2016.

[61] T. Wimalajeewa and P. K. Varshney, "OMP based joint sparsity pattern recovery under communication constraints," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5059–5072, 2014.

[62] D. Eiwen, G. Tauböck, F. Hlawatsch, H. Rauhut, and N. Czink, "Multichannel-compressive estimation of doubly selective channels in MIMO-OFDM systems: Exploiting and enhancing joint sparsity," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 3082–3085.

[63] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

[64] J. S. Yedidia, "Message-passing algorithms for inference and optimization," *Journal of Statistical Physics*, vol. 145, no. 4, pp. 860–890, 2011.

[65] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[66] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, ser. Communications and Signal Processing. McGraw-Hill, 1991.

[67] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[68] P. Maechler, C. Studer, D. E. Bellasi, A. Maleki, A. Burg, N. Felber, H. Kaeslin, and R. G. Baraniuk, "VLSI design of approximate message passing for signal restoration and compressive sensing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 3, pp. 579–590, 2012.

[69] G. Hannak, M. Mayer, G. Matz, and N. Goertz, "Bayesian QAM demodulation and activity detection for multiuser communication systems," in *2016 IEEE International Conference on Communications Workshops (ICC)*. IEEE, 2016, pp. 596–601.

[70] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 454–465, 2011.

[71] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: II. analysis and validation," in *2010 IEEE Workshop on Information Theory*, Jan. 2010, pp. 1–5.

[72] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[73] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[74] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6920–6941, 2011.

[75] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing, "Greedy algorithms for joint sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1694–1704, 2014.

[76] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method," in *SPIE Optical Engineering + Applications*. International Society for Optics and Photonics, 2013, pp. 88 580R–88 580R.

[77] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[78] J. Kim, W. Chang, B. Jung, D. Baron, and J. C. Ye, "Belief propagation for joint sparse recovery," *arXiv preprint arXiv:1102.3289*, 2011.

[79] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[80] G. Hannak, M. Mayer, G. Matz, and N. Goertz, "An approach to complex bayesian-optimal approximate message passing," *arXiv preprint arXiv:1511.08238*, 2015.

[81] Y. Lu and W. Dai, "Independent versus repeated measurements: A performance quantification via state evolution," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4653–4657.

[82] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[83] J. Zhu, D. Baron, and F. Krzakala, "Performance limits for noisy multimeasurement vector problems," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2444–2454, 2017.

[84] M. Mézard and A. Montanari, *Information, Physics, and Computation*. Oxford University Press, 2009.

[85] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 08, p. P08009, 2012.

[86] ——, "Statistical-physics-based reconstruction in compressed sensing," *Physical Review X*, vol. 2, no. 2, p. 021005, 2012.

[87] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 665–669.

[88] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Transactions on Information Theory*, vol. 48, no. 11, pp. 2888–2910, 2002.

[89] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[90] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity-achieving sparse superposition codes," *arXiv preprint arXiv:1503.08040*, 2015.

[91] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4719–4734, 2011.

[92] S. Jalali and A. Maleki, "From compression to compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 352–385, 2016.

[93] S. Birgmeier, *Utilization of correlation between signal components for compressed-sensing recovery*, 2015, wien, Techn. Univ., Dipl.-Arb., 2015.

[94] J. Ender, "On compressive sensing applied to radar," *Signal Processing*, vol. 90, no. 5, pp. 1402–1414, 2010.

[95] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2275–2284, 2009.

[96] Q. Huang, L. Qu, B. Wu, and G. Fang, "UWB through-wall imaging based on compressive sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1408–1415, 2010.

[97] Y. Yu, A. P. Petropulu, and H. V. Poor, "Measurement matrix design for compressive sensing–based MIMO radar," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5338–5352, 2011.

[98] A. Massa, P. Rocca, and G. Oliveri, "Compressive sensing in electromagnetics - a review," *IEEE Antennas and Propagation Magazine*, vol. 57, no. 1, pp. 224–238, 2015.

[99] D. Schneider, "New camera chip captures only what it needs," *IEEE Spectrum*, vol. 50, no. 3, pp. 13–14, March 2013.

[100] F. Maia, A. MacDowell, S. Marchesini, H. A. Padmore, D. Y. Parkinson, J. Pien, A. Schirotzek, and C. Yang, "Compressive phase contrast tomography," *arXiv preprint arXiv:1009.1380*, 2010.

[101] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[102] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[103] Y. Zhang, B. S. Peterson, G. Ji, and Z. Dong, "Energy preserved sampling for compressed sensing MRI," *Computational and Mathematical Methods in Medicine*, vol. 2014, 2014.

[104] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, and J. Yang, "Exponential wavelet iterative shrinkage thresholding algorithm for compressed sensing magnetic resonance imaging," *Information Sciences*, vol. 322, pp. 115–132, 2015.

[105] L. Anitori, M. Otten, W. Van Rossum, A. Maleki, and R. Baraniuk, "Compressive CFAR radar detection," in *IEEE Radar Conference (RADAR)*. IEEE, 2012, pp. 0320–0325.

[106] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4290–4308, 2013.

[107] S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid generalized approximate message passing with applications to structured sparsity," in *IEEE International Symposium on Information Theory*. IEEE, 2012, pp. 1236–1240.

[108] R. Want, "An introduction to RFID technology," *IEEE Pervasive Computing*, vol. 5, no. 1, pp. 25–33, 2006.

[109] D. M. Dobkin, *The RF in RFID: UHF RFID in Practice*. Newnes, 2012.

[110] C. Boyer and S. Roy, "Backscatter communication and RFID: Coding, energy, and MIMO analysis," *IEEE Transactions on Communications*, vol. 62, no. 3, pp. 770–785, 2014.

[111] S. Verdú, *Multiuser Detection.* Cambridge (UK): Cambridge University Press, 1998.

[112] A. Bartoli, M. Dohler, J. Hernández-Serrano, A. Kountouris, and D. Barthel, "Low-power low-rate goes long-range: The case for secure and cooperative machine-to-machine communications," in *Networking 2011 Workshops.* Springer, 2011, pp. 219–230.

[113] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 389–400, 2013.

[114] H. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," in *8th International Symposium on Wireless Communication Systems (ISWCS)*, Nov. 2011, pp. 291–295.

[115] H. Schepker, C. Bockelmann, and A. Dekorsy, "Coping with CDMA asynchronicity in compressive sensing multi-user detection," in *IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[116] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Communications Letters*, vol. 16, no. 7, pp. 972–974, July 2012.

[117] J. G. Proakis, M. Salehi, N. Zhou, and X. Li, *Communication Systems Engineering.* Prentice Hall New Jersey, 1994, vol. 2.

[118] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," *Technical University of Denmark*, vol. 7, p. 15, 2008.

[119] M. Raphan and E. P. Simoncelli, "Empirical Bayes least squares estimation without an explicit prior," *NYU Courant Inst. Tech. Report*, 2007.

[120] ——, "Least squares estimation without priors or supervision," *Neural Computation*, vol. 23, no. 2, pp. 374–420, 2011.