**Aalborg Universitet**

AALBORG UNIVERSITY
DENMARK

# Sparse Variational Bayesian SAGE Algorithm With Application to the Estimation of Multipath Wireless Channels

Shutin, Dmitriy; Fleury, Bernard Henri

*Published in:*
I E E E Transactions on Signal Processing

*DOI (link to publication from Publisher):*
10.1109/TSP.2011.2140106

*Publication date:*
2011

*Document Version*
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

*Citation for published version (APA):*
Shutin, D., & Fleury, B. H. (2011). Sparse Variational Bayesian SAGE Algorithm With Application to the Estimation of Multipath Wireless Channels. *I E E E Transactions on Signal Processing*, *59*(8), 3609-3623. https://doi.org/10.1109/TSP.2011.2140106

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

**Take down policy**
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

# Sparse Variational Bayesian SAGE Algorithm With Application to the Estimation of Multipath Wireless Channels

Dmitriy Shutin, *Member, IEEE*, and Bernard H. Fleury, *Senior Member, IEEE*

***Abstract*—In this paper, we develop a sparse variational Bayesian (VB) extension of the space-alternating generalized expectation-maximization (SAGE) algorithm for the high resolution estimation of the parameters of relevant multipath components in the response of frequency and spatially selective wireless channels. The application context of the algorithm considered in this contribution is parameter estimation from channel sounding measurements for radio channel modeling purpose. The new sparse VB-SAGE algorithm extends the classical SAGE algorithm in two respects: i) by monotonically minimizing the variational free energy, distributions of the multipath component parameters can be obtained instead of parameter point estimates and ii) the estimation of the number of relevant multipath components and the estimation of the component parameters are implemented jointly. The sparsity is achieved by defining parametric sparsity priors for the weights of the multipath components. We revisit the Gaussian sparsity priors within the sparse VB-SAGE framework and extend the results by considering Laplace priors. The structure of the VB-SAGE algorithm allows for an analytical stability analysis of the update expression for the sparsity parameters. This analysis leads to fast, computationally simple, yet powerful, adaptive selection criteria applied to the single multipath component considered at each iteration. The selection criteria are adjusted on a per-component-SNR basis to better account for model mismatches, e.g., diffuse scattering, calibration and discretization errors, allowing for a robust extraction of the relevant multipath components. The performance of the sparse VB-SAGE algorithm and its advantages over conventional channel estimation methods are demonstrated in synthetic single-input–multiple-output (SIMO) time-invariant channels. The algorithm is also applied to real measurement data in a multiple-input–multiple-output (MIMO) time-invariant context.**

***Index Terms*—Expectation-maximization algorithm, MIMO, multipath channels, SAGE algorithm, variational Bayesian methods.**

## I. Introduction

In modeling real world data, proper model selection plays a pivotal role. When applying high resolution algorithms to the estimation of wireless multipath channels from multidimensional channel measurements, an accurate determination of the number of dominant multipath components is required in order to reproduce the channel behavior in a realistic manner—an essential driving mechanisms for the design and development of next generation multiple-input–multiple-output (MIMO)-capable wireless communication and localization systems. Consider for simplicity a single-input–multiple-output (SIMO) wireless channel,[1] e.g., an uplink channel with a base station equipped with multiple antennas. The received signal vector $\boldsymbol{z}(t)$ made of the signals at the outputs of these antennas can be represented as a superposition of an unknown number $L$ of multipath components $w_l \boldsymbol{s}(t, \boldsymbol{\theta}_l)$ contaminated by additive noise $\boldsymbol{\xi}(t)$ [1]

$$\boldsymbol{z}(t) = \sum_{l=1}^{L} w_l \boldsymbol{s}(t, \boldsymbol{\theta}_l) + \boldsymbol{\xi}(t). \tag{1}$$

In (1) $w_l$ is the multipath weights and $\boldsymbol{s}(t, \boldsymbol{\theta}_l)$ is the received version of the transmitted signal modified according to the dispersion parameter vector $\boldsymbol{\theta}_l$ of the $l$th propagation path.[2] Classical parameter estimation [2]–[5] deals with the estimation of the multipath components, i.e., $w_l$ and $\boldsymbol{\theta}_l$, while the estimation of the number $L$ of these components is the object of model order

Manuscript received March 23, 2010; revised September 17, 2010 and February 24, 2011; accepted March 16, 2011. Date of publication April 07, 2011; date of current version July 13, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Coates. This work was supported in part by the Austrian Science Fund (FWF) by Grant NFN SISE (S106), by the European Commission within the ICT-216715 FP7 Network of Excellence in Wireless Communications (NEWCOM++), by the project ICT-217033 Wireless Hybrid Enhanced Mobile Radio Estimators (WHERE), and by the project ICT-248894 (WHERE-2).

D. Shutin is with the Department of Electrical Engineering, Princeton University, B311 E-QUAD, Princeton NJ 08544 USA (e-mail: dshutin@princeton.edu).

B. H. Fleury is with the Section Navigation and Communications, Department of Electronic Systems, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: bfl@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2140106

[1]The proposed method can be easily extended to MIMO time-variant channels with stationary propagation constellation. With minor modifications the polarization aspects can be included as well. This extension merely leads to a more complicated signal model, including for instance more dispersion parameters, without adding any new aspect relevant to the understanding of the new proposed concepts and methods. The scenario considering a SIMO channel seems a sensible compromise between complexity of the model underlying the theoretical analyses and an interesting application in which the proposed method can be demonstrated. However, in the experimental section we consider the estimation of a MIMO channel.

[2]We mean as dispersion parameters of the waves propagating from the transmitter side to the receiver site—and, by generalization, of the multipath components in the resulting channel response—their relative delay, direction of departure, direction of arrival, and Doppler frequency. The parameter $\boldsymbol{\theta}_l$ includes all these parameters or a subset of them depending on the transmitter and receiver configurations.

1053-587X/$26.00 © 2011 IEEE

selection [6]–[9]. Despite its obvious simplicity, (1) provides an oversimplified description of reality: it adequately accounts for specular-like propagation paths only. Components originating from diffuse scattering inevitably present in measured channel responses are not rendered appropriately in (1). More specifically, a very large number of specular components $L$ is needed to represent such diffuse components. Further effects leading to model mismatch are errors in calibration of the response of the transceiver or measurement equipment that cause inaccuracies in the description of $\boldsymbol{s}(t,\boldsymbol{\theta}_l)$, as well as the discrete-time approximation to (1), typically arising when model parameters are estimated using numerical optimization techniques. All these effects have a significant impact on the performance of both the parameter estimation algorithms and the model order selection schemes derived based on (1). Experimental evidence shows that if the model order selection scheme is not carefully designed, the above model mismatch will lead to an overestimation of the number of relevant multipath components. Fictive components without any physical meaning will be introduced and their parameters estimated. Hence, radio channel estimators combining model order (component number) selection and component parameter estimation that are robust against model mismatch are needed here.

Bayesian methods are promising candidates for such robust methods. For a fixed model order $L$ the classical maximum likelihood (ML) approach to the estimation of dispersion parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L\}$ and gains $\boldsymbol{w} = \{w_1,\ldots,w_L\}$ in (1) involves maximization of the multidimensional parameter likelihood $p(\boldsymbol{z}|\boldsymbol{\Theta},\boldsymbol{w})$ given the measurement $\boldsymbol{z}$. Although efficient algorithms exist to solve this optimization problem [2], [3], [10], standard ML algorithms require a fixed number of components $L$ and typically do not employ any likelihood penalization to compensate for overfitting. Bayesian techniques can compensate for this through the use of a prior $p(\boldsymbol{\Theta},\boldsymbol{w})$, which effectively imposes constrains on the model parameters. The model fit (i.e., the value of the likelihood) can be traded for the model complexity [i.e., number of components in (1)] through the likelihood penalization. Likelihood penalization lies in the heart of celebrated information-theoretic model order selection criteria, such as minimum description length (MDL), Bayesian information criterion (BIC), as well as their variants [7]–[9].

Imposing constraints on the model parameters is a key to sparse signal modeling [11]–[16]. In Bayesian sparsity approach [11], [13], [14], [17] the gains $\boldsymbol{w}$ are constrained using a parametric prior $p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_l p(w_l|\alpha_l)$, where $p(w_l|\alpha_l)$ is a circularly symmetric probability density function (pdf), with the prior parameter $\alpha_l$—also called sparsity parameter—being inversely proportional to the width of the pdf. Such form of the prior allows for controlling the contribution of each basis associated with the weight $w_l$ through the sparsity parameter $\alpha_l$: a large value of $\alpha_l$ will drive the corresponding weight $w_l$ to zero, thus realizing a sparse estimator. The sparsity parameters are found as the maximizers of $p(\boldsymbol{z}|\boldsymbol{\alpha})$, which is also known as a type-II likelihood function or model evidence [13], [14], [18] and the corresponding estimation approach is known as the evidence procedure (EP) [14].

In general, evaluating $p(\boldsymbol{z}|\boldsymbol{\alpha})$ is difficult. This, however, can be done analytically [11], [13], [14], [17] in the special case of linear models[3] $p(\boldsymbol{z}|\boldsymbol{\Theta},\boldsymbol{w})$ with both model distribution and sparsity prior being Gaussian. This choice of the prior pdf corresponds to the $\ell_2$-type of parameter constraints. Moreover, it can be shown [19] that in the Gaussian prior case the maximum of the model evidence $p(\boldsymbol{z}|\boldsymbol{\alpha})$ coincides with the Bayesian interpretation of the normalized ML model order selection [7] and reduces to the BIC as the number of measurement samples grows. Therefore, the EP allows for joint model order selection and parameter estimation. This approach was investigated in [19] within the context of wireless channels; however, [19] considers the estimation of multipath gains only, thus bypassing the estimation of the dispersion parameters in (1). Recently, several investigations have been dedicated to study the $\ell_1$-type of parameter penalties [12], [15], [16], [20], [21], which, in the Bayesian sparsity framework, is equivalent to choosing $p(w_l|\alpha_l)$ as a Laplace prior for $l = 1,\ldots,L$. Compared to Gaussian priors, such form of constraints leads to sparser models [13], [15], [22], [23]. The $\ell_1$-type of penalties significantly limits the analytical study of the algorithm; nonetheless for models linear in their parameters different efficient numerical techniques have been developed [15], [24], [25]. The extension of the Bayesian sparsity methods with Laplace priors applied to the estimation of multipath wireless channels has not been explored yet, mainly due to the nonlinearity of the channel model in $\boldsymbol{\Theta}$. This can be circumvented by using virtual channel models [16], [21], which is equivalent to a sampling or gridding of the dispersion parameters $\boldsymbol{\Theta}$ at the Nyquist rate [16]. The algorithm then estimates the coefficients on the grid using sparsity techniques [12], [16], [21]. This approach, however, does not provide high resolution estimates of the multipath parameters. Although it is very effective in capturing channel energy, recent investigations [26] demonstrate that this approach inevitably leads to a mismatch between the true channel sparsity and the estimated sparsity; more specifically, even when fine quantization of $\boldsymbol{\Theta}$ is used, the number of virtual multipath components will always exceed the true number of multipath components; in that respect the channel estimates derived based on virtual models are not appropriate when the goal is to extract physical multipath components. In this paper we aim to demonstrate that the superresolution property should not be sacrificed to the linearity of the estimation problem. We achieve this by: i) casting a super-resolution SAGE algorithm for multipath parameter estimation [3] in a Bayesian framework, and treating the entries in $\boldsymbol{\Theta}$ as random variables whose pdfs are to be estimated and ii) combining this estimation scheme with the Bayesian sparsity techniques, as aforementioned, i.e., using multiple sparsity parameters $\alpha_l$ to control the model sparsity on a per-component basis. Moreover, as we will show, our analysis also allows for defining ways to reduce the impact of estimation artifacts due to the basis mismatches through a detailed analysis of the estimation expressions for the sparsity parameters.

Our main contribution in this paper is twofold. First, in order to realize Bayesian sparse estimation and to overcome the computational difficulties due to the nonlinearity of the

[3]In our context this corresponds to assuming $\boldsymbol{\theta}_l$ as known or fixed, and thus $\boldsymbol{s}(\boldsymbol{\theta}_l) \equiv \bar{\boldsymbol{s}}$.

channel model, we propose a new variational Bayesian (VB) [27] extension of the space-alternating generalized expectation-maximization (SAGE) algorithm for multipath parameter estimation [3], [28]. We coin this extension the variational Bayesian SAGE (VB-SAGE). In contrast to the SAGE algorithm, the VB-SAGE algorithm estimates the posterior pdfs of the model parameters by approximating the true posterior pdf $p(\boldsymbol{\Theta}, \boldsymbol{w}, \boldsymbol{\alpha}|\boldsymbol{z})$ with a proxy pdf $q(\boldsymbol{\Theta}, \boldsymbol{w}, \boldsymbol{\alpha})$ such as to minimize the variational free energy [27]. Similar to the original SAGE algorithm [28], the VB-SAGE algorithm relies on the concept of the admissible hidden data—an analog of the complete data in the EM framework—to optimize at each iteration the variational free energy with respect to the pdfs of the parameters of one component only. We demonstrate that the monotonicity property of the VB-SAGE algorithm guarantees that such optimization strategy necessarily minimizes the variational free energy. Such optimization strategy makes the estimation of the parameters in $\boldsymbol{\Theta}$ a tractable optimization problem due to the reduced dimensionality of the resulting objective functions. Second, we demonstrate that the admissible hidden data also permits a detailed analytical study of the sparsity parameters $\boldsymbol{\alpha}$, which leads to selection criteria applied individually to the multipath component updated at each iteration. On the one hand, these selection criteria allow for a fast implementation of the sparse channel estimator; on the other hand, they are easy to interpret and can be adjusted to compensate for model mismatch due to, e.g., calibration and discretization errors. Thus, the sparse VB-SAGE algorithm jointly implements the estimation of the number of relevant multipath components and the estimation of the posterior pdfs of the component parameters. We revisit and extend the Gaussian prior case, and present new results for Laplace sparsity priors within the framework of the VB-SAGE algorithm. It should also be mentioned that the performed analysis of the sparsity parameters $\boldsymbol{\alpha}$ is equally valid for the problem of sparse estimation of virtual channel models [16] with the VB-SAGE algorithm. However, the application of the sparse VB-SAGE algorithm to the estimation of virtual channel models is outside the scope of the paper.

The paper is organized as follows: In Section II we introduce the signal model; Section III addresses the derivation of the VB-SAGE algorithm for the multipath parameter estimation, followed by the analysis of the sparsity priors for model order selection discussed in Section IV; in Section V several practical issues, e.g., algorithm initialization, are discussed; in Section VI estimation results obtained from synthetic and measured data are presented; finally, we conclude the paper in Section VII.

Through the paper we shall make use of the following notation. Vectors are represented as boldface lowercase letters, e.g., $\boldsymbol{x}$, and matrices as boldface uppercase letters, e.g., $\boldsymbol{X}$. For vectors and matrices $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and Hermitian transpose, respectively. Sets are represented as calligraphic uppercase letters, e.g., $\mathcal{S}$. We use $\mathcal{I}$ to denote an index set, i.e., $\mathcal{I} = \{1, \ldots, L\}$. The assumed number of elements in $\mathcal{I}$ is $L$, unless stated otherwise. We will write $\boldsymbol{x}_{\mathcal{I}} \equiv \bigcup_{l=1}^{L} \{\boldsymbol{x}_l\}$ as a shorthand notation for a list of variables $\boldsymbol{x}_l$ with indices $l \in \mathcal{I}$. When $\mathcal{S}$ is a set and $\boldsymbol{x} \in \mathcal{S}$, then $\overline{\mathcal{S}(\boldsymbol{x})} = \mathcal{S} \setminus \{\boldsymbol{x}\}$ is the complement of $\{\boldsymbol{x}\}$ in $\mathcal{S}$. Similarly, $\overline{\mathcal{I}(l)} = \mathcal{I} \setminus \{l\}$ and $\overline{\mathcal{S}(\mathcal{M})} = \mathcal{S} \setminus \mathcal{M}$. Two types of proportionality are used: $x \propto y$ denotes $x = \alpha y$; $x \propto^e y$ denotes $\mathrm{e}^x = \mathrm{e}^{\beta}\mathrm{e}^y$ and thus $x = \beta + y$ for some arbitrary constants $\alpha$ and $\beta$. An estimate of a random variable $\boldsymbol{x}$ is denoted as $\widehat{\boldsymbol{x}}$. We use $\mathbb{E}_{q(\boldsymbol{x})}\{f(\boldsymbol{x})\}$ to denote the expectation of a function $f(\boldsymbol{x})$ with respect to a probability density $q(\boldsymbol{x})$; similarly, $\mathbb{E}_{q(\mathcal{M})}\{f(\boldsymbol{x})\}$ denotes the expectation with respect to the joint probability density $q(\mathcal{M})$ of the random variables in the set $\mathcal{M}$. Finally, $\mathrm{CN}(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{B})$ denotes a multivariate complex Gaussian pdf with a mean $\boldsymbol{a}$ and a covariance matrix $\boldsymbol{B}$; $\mathrm{Ga}(x; a, b)$ denotes a gamma pdf with parameters $a$ and $b$.

## II. Signal Model

Channel sounding is an instrumental method for the design of accurate and realistic radio channel models. Channel sounding is usually performed by sending a specific sounding sequence $u(t)$ through the channel and observing the response $z(t)$ at the receiving side. The received signal $z(t)$ is then used to estimate the channel impulse response (CIR) or its parameters when a parametric model of the response is specified. Consider now a SIMO channel model and time-domain channel sounding. The sounding signal $u(t)$ consists of $N_{\mathrm{u}}$ periodically repeated burst waveforms $b(t)$, i.e., $u(t) = \sum_{i=0}^{N_{\mathrm{u}}-1} b(t - iT_{\mathrm{f}})$, where $b(t)$ has duration $T_{\mathrm{b}} \leq T_{\mathrm{f}}$ and is formed as $b(t) = \sum_{m=0}^{M-1} b_m p(t - mT_{\mathrm{p}})$. The known sounding sequence $b_0 \ldots b_{M-1}$ consists of $M$ chips and $p(t)$ is the shaping pulse of duration $T_{\mathrm{p}}$, with $MT_{\mathrm{p}} = T_{\mathrm{b}}$. We assume that the signal vector $\boldsymbol{z}(t)$ has been received/measured with an antenna array consisting of $M_{\mathrm{r}}$ sensors located at positions $\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{M_{\mathrm{r}}-1} \in \mathbb{R}^2$ with respect to an arbitrary reference coordinate system. The signal originating from the $l$th propagation path is an altered version of the original transmitted signal $u(t)$ weighted by a complex gain $w_l$. The alteration process is described by a (nonlinear) mapping $u(t) \mapsto \boldsymbol{s}(t, \boldsymbol{\theta}_l)$, where $\boldsymbol{\theta}_l$ is the vector of dispersion parameters, e.g., relative delay, azimuth and elevation angles of arrival. The nonlinear mapping $u(t) \mapsto \boldsymbol{s}(t, \boldsymbol{\theta}_l)$ includes the system effects, e.g., the transmitter and receiver RF/IF filters and the responses of the transmit and receive arrays. In the sequel we try to abstract from the concrete channel structure where it is possible and keep the model in its most general form. Additive noise $\boldsymbol{\xi}(t)$ is assumed to be a zero-mean spatially white and temporally wide-sense stationary Gaussian process, i.e., $E\{\xi_k(t)\xi_k^*(t+\tau)\} = R_\xi(\tau)$, and $E\{\xi_m(t)\xi_k^*(t+\tau)\} = 0$, $0 \leq k, m \leq M_{\mathrm{r}} - 1$, $k \neq m$. In our framework we assume that $R_\xi(\tau)$ is known.[4] In practice $\boldsymbol{z}(t)$ is low-pass filtered and sampled with the sampling period $T_{\mathrm{s}}$, resulting in $M_{\mathrm{r}}$ $N$-tuples with $N$ being the number of output samples per sensor. By stacking the sampled outputs of the $M_{\mathrm{r}}$ sensors in one vector $\boldsymbol{z}$, (1) can be rewritten as

$$\boldsymbol{z} = \sum_{l=1}^{L} w_l \boldsymbol{s}(\boldsymbol{\theta}_l) + \boldsymbol{\xi} \tag{2}$$

where we define $\boldsymbol{s}(\boldsymbol{\theta}_l) = [\boldsymbol{s}_0(\boldsymbol{\theta}_l)^T, \ldots, \boldsymbol{s}_{M_{\mathrm{r}}-1}(\boldsymbol{\theta}_l)^T]^T$, $\boldsymbol{\xi} = [\boldsymbol{\xi}_0^T, \ldots, \boldsymbol{\xi}_{M_{\mathrm{r}}-1}^T]^T$, with $\boldsymbol{s}_p(\boldsymbol{\theta}_l) = [s_p(0, \boldsymbol{\theta}_l), \ldots, s_p((N-1)T_{\mathrm{s}}, \boldsymbol{\theta}_l)]^T$, and $\boldsymbol{\xi}_p = [\xi_p(0), \ldots, \xi_p((N-1)T_{\mathrm{s}})]^T$, $p = 0, \ldots, M_{\mathrm{r}} - 1$. Finally, we define $\boldsymbol{\Omega} = \{w_1, \boldsymbol{\theta}_1, \ldots, w_L, \boldsymbol{\theta}_L\}$.

[4] Although it is possible to reformulate the algorithm to estimate the noise covariance [14], [29], we will leave this aspect outside the scope of this work.

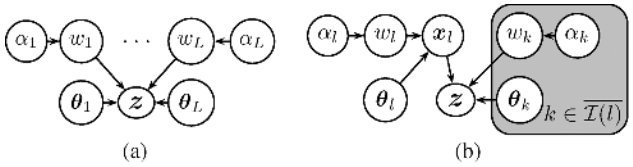


Fig. 1. (a) Graphical model representing (2) with $L$ components. (b) Extended model with the admissible hidden data $\boldsymbol{x}_l$.

The probabilistic graph depicted in Fig. 1(a) encodes the dependencies between the parameters and the observation vector in the model (2). As visualized in the graph structure, the joint pdf of the probabilistic model can be factored as $p(\boldsymbol{z}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = p(\boldsymbol{z}|\boldsymbol{\Omega})p(\boldsymbol{\Omega}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^T$ is the vector containing the model sparsity parameters. Let us now specify the statistical model behind the involved variables.

Under the Gaussian noise assumption, $p(\boldsymbol{z}|\boldsymbol{\Omega}) = \mathrm{CN}(\boldsymbol{z}; \sum_{l=1}^{L} w_l \boldsymbol{s}(\boldsymbol{\theta}_l), \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = E\{\boldsymbol{\xi}\boldsymbol{\xi}^H\}$. The second term $p(\boldsymbol{\Omega}|\boldsymbol{\alpha})$ is the parameter prior. We assume that $p(\boldsymbol{\Omega}|\boldsymbol{\alpha}) = \prod_{l=1}^{L} p(w_l|\alpha_l)p(\boldsymbol{\theta}_l)$, where $p(w_l|\alpha_l)$ is the sparsity prior for the $l$th component. The purpose of the sparsity prior is, on the one hand, to constrain the gains $w_1, \ldots, w_L$ of the components, and thus implement sparsification/model order selection, and, on the other hand, to control this constraint through the sparsity parameters $\boldsymbol{\alpha}$. We will study two choices for $p(w_l|\alpha_l)$: i) a Gaussian prior, and ii) a Laplace prior. In both cases the prior pdfs are complex circularly symmetric, with the nonnegative hyperparameter $\alpha_l$ inversely proportional to their width. Thus, large values of $\alpha_l$ will render the contribution of the component $w_l \boldsymbol{s}(\boldsymbol{\theta}_l)$ "irrelevant" since the corresponding prior over $w_l$ will then be concentrated at the origin. The choice of the prior $p(\boldsymbol{\theta}_l)$ is arbitrary; however, it must reflect the underlying physics and restrictions of the measurement equipment; a non-informative prior can also be used. The prior $p(\alpha_l)$, also called the hyperprior of the $l$th component, is selected as a gamma pdf $\mathrm{Ga}(\alpha_l; a_l, b_l) = \frac{b_l^{a_l}}{\Gamma(a_l)} \alpha_l^{a_l - 1} \exp(-b_l \alpha_l)$. Practically we set $a_l = b_l = 10^{-7}$ for all components to render their hyperprior non-informative [13], [14]. Such formulation of a hyperprior pdf $p(\boldsymbol{\alpha}) = \prod_l p(\alpha_l)$ is related to automatic relevance determination [18], [30].

## III. Parameter Estimation Framework

Direct evaluation of $p(\boldsymbol{z}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ or of the posterior $p(\boldsymbol{\Omega}, \boldsymbol{\alpha}|\boldsymbol{z})$ for performing inference of the unknown parameters $\{\boldsymbol{\Omega}, \boldsymbol{\alpha}\}$ is a nontrivial task. Two main reasons for this are the nonlinearity of the model (1) and the statistical dependence of multipath component parameters when $\boldsymbol{z}$ is observed.[5] Approximative techniques might significantly ease the model fitting step. In our work we resort to the variational Bayesian inference framework. The variational Bayesian inference generalizes the classical EM algorithm [27] and provides a tool for estimating distributions of $\{\boldsymbol{\Omega}, \boldsymbol{\alpha}\}$. Essentially, variational methods approximate the posterior pdf of interest with a simpler pdf (by, e.g., neglecting some statistical dependencies between random variables) such that the Kullback-Leibler divergence between the former pdf and the latter is minimized.

When estimating parameters using the SAGE algorithm [3], [28], the concept of complete data in the EM algorithm is replaced by that of admissible hidden data. The purpose of the admissible hidden data is to make the update procedure for only a subset $\boldsymbol{\Omega}_{\mathrm{sub}} \subset \boldsymbol{\Omega}$ a tractable optimization problem. For the variable $\boldsymbol{x}_{\mathrm{sub}}$ to be an admissible hidden data with respect to $\boldsymbol{\Omega}_{\mathrm{sub}}$ the following factorization must be satisfied: $p(\boldsymbol{z}, \boldsymbol{x}_{\mathrm{sub}}, \boldsymbol{\Omega}) = p(\boldsymbol{z}|\boldsymbol{x}_{\mathrm{sub}}, \{\boldsymbol{\Omega} \setminus \boldsymbol{\Omega}_{\mathrm{sub}}\})p(\boldsymbol{x}_{\mathrm{sub}}, \boldsymbol{\Omega})$ [28]. The fact that $\boldsymbol{x}_{\mathrm{sub}}$ is an admissible hidden data guarantees that the likelihood of the new parameter update $\boldsymbol{\Omega}'$ (obtained by replacing the updated parameter subset $\boldsymbol{\Omega}'_{\mathrm{sub}}$ in the overall parameter set $\boldsymbol{\Omega}$) cannot be smaller than the likelihood prior to the update [28]. This property is referred to as the monotonicity property. The concept of admissible hidden data can be exploited within the variational framework as well. As we will show later, this similarly leads to an iterative algorithm—we call it the VB-SAGE algorithm—that still exhibits the monotonicity property in terms of the variational free energy [27].

Consider for a specific component $l$ the new variable

$$\boldsymbol{x}_l = w_l \boldsymbol{s}(\boldsymbol{\theta}_l) + \boldsymbol{\xi}_l \tag{3}$$

which can be conceived as a received signal associated with the $l$th propagation path. The additive noise component $\boldsymbol{\xi}_l$ in (3) is obtained by arbitrarily decomposing the total noise $\boldsymbol{\xi}$ such that $\boldsymbol{\Sigma}_l = E\{\boldsymbol{\xi}_l \boldsymbol{\xi}_l^H\} = \beta_l \boldsymbol{\Sigma}$ and $0 \leq \beta_l \leq 1$. We define $\boldsymbol{\Delta}_l = (1 - \beta_l)\boldsymbol{\Sigma}$ to be the part of the total additive noise that is not associated with the $l$th component. Thus, $\boldsymbol{\Delta}_l + \boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}$. Consider now the modified graph in Fig. 1(b) that accounts for $\boldsymbol{x}_l$. It is straightforward to show that $\boldsymbol{x}_l$ is an admissible hidden data with respect to the subset $\{w_l, \boldsymbol{\theta}_l, \alpha_l\}$. Since we are interested in estimating all $L$ components, we can formulate the estimation algorithm as a succession of $L$ estimations of $\{\boldsymbol{\theta}_l, w_l, \alpha_l\}$ with respect to $\boldsymbol{x}_l$, $l = 1, \ldots, L$, assuming that $\{w_k, \boldsymbol{\theta}_k, \alpha_k\}$, $k \in \overline{\mathcal{I}(l)}$, are known and fixed. According to the extended graph in Fig. 1(b), the joint pdf $p(\boldsymbol{z}, \boldsymbol{x}_l, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ now factors as

$$p\left(\boldsymbol{z}|\boldsymbol{x}_l, \boldsymbol{\theta}_{\overline{\mathcal{I}(l)}}, w_{\overline{\mathcal{I}(l)}}\right) \prod_{k \in \overline{\mathcal{I}(l)}} p(\boldsymbol{\theta}_k)p(w_k|\alpha_k)p(\alpha_k) \times p(\boldsymbol{x}_l|w_l, \boldsymbol{\theta}_l)p(w_l|\alpha_l)p(\boldsymbol{\theta}_l)p(\alpha_l) \tag{4}$$

where

$$p\left(\boldsymbol{z}|\boldsymbol{x}_l, \boldsymbol{\theta}_{\overline{\mathcal{I}(l)}}, w_{\overline{\mathcal{I}(l)}}\right) = \mathrm{CN}\left(\boldsymbol{z}; \boldsymbol{x}_l + \sum_{k \in \overline{\mathcal{I}(l)}} w_k \boldsymbol{s}(\boldsymbol{\theta}_k), \boldsymbol{\Delta}_l\right) \tag{5}$$

and $p(\boldsymbol{x}_l|w_l, \boldsymbol{\theta}_l) = \mathrm{CN}(\boldsymbol{x}_l; w_l \boldsymbol{s}(\boldsymbol{\theta}_l), \boldsymbol{\Sigma}_l)$.

### A. Variational Bayesian Inference of Signal Parameters

Variational Bayesian inference [27] is a family of techniques that exploit analytical approximations of the posterior pdf of interest, i.e., $p(\boldsymbol{\Omega}, \boldsymbol{\alpha}|\boldsymbol{z})$, using a simpler proxy pdf $q(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. The latter pdf is estimated as a minimizer of the variational free energy $\mathcal{F}(q(\boldsymbol{\Omega}, \boldsymbol{\alpha})||p(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{z}))$ [27], which is formally equivalent to the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Omega}, \boldsymbol{\alpha})||p(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{z}))$ between the proxy pdf and the true joint pdf. The admissible

[5]Such graph structure is also referred to as a V-structure [31], which leads to the conditional dependence of the parent variables when the corresponding child variable is observed.

hidden data, used in the SAGE algorithm to facilitate the maximization of the parameter likelihood, can also be used within the variational inference framework to ease the minimization of the variational free energy. Such algorithm we term a VB-SAGE algorithm.

Essentially, the VB-SAGE algorithm approximates $p(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{x}_l, \boldsymbol{z})$ with a variational proxy pdf

$$q(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{x}_l) = q(\boldsymbol{x}_l) \prod_{k=1}^{L} q(w_k) q(\boldsymbol{\theta}_k) q(\alpha_k) \tag{6}$$

by minimizing the free energy with respect to the parameter of the $l$th component only, and cycling through all $L$ components in a "round-robin" fashion. The monotonicity property of the VB-SAGE algorithm (see Appendix A) ensures that such sequential optimization necessarily decreases the free energy $\mathcal{F}(q(\boldsymbol{\Omega}, \boldsymbol{\alpha}) || p(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{z}))$.

It is straightforward to show that with the factorization (6) the estimation of any factor $q(a)$, $a \in \{w_{\mathcal{I}}, \boldsymbol{\theta}_{\mathcal{I}}, \alpha_{\mathcal{I}}, \boldsymbol{x}_l\}$, requires the Markov blanket $\mathcal{MB}(a)$ [31] of $a$ to be known.[6] Define now

$$\tilde{p}(a) \propto \exp\left(\mathbb{E}_{q(\mathcal{MB}(a))}\left\{\log p\left(a | \mathcal{MB}(a)\right)\right\}\right). \tag{7}$$

The unconstrained solution for $q(a)$ that minimizes the corresponding free energy is then simply found as $q(a) = \tilde{p}(a)$. Clearly, an unconstrained solution is preferred. However, we might have to constrain $q(a)$ to belong to some class $\mathcal{Q}(a)$ of pdfs in order to make the optimization tractable. In this case the approximate solution is obtained by solving

$$q(a) = \underset{q^*(a) \in \mathcal{Q}(a)}{\arg\min}\ \mathrm{D}_{\mathrm{KL}}\left(q^*(a) || \tilde{p}(a)\right). \tag{8}$$

In the case of $\boldsymbol{x}_l$ it is straightforward to show that $\tilde{p}(\boldsymbol{x}_l)$ is quadratic in $\boldsymbol{x}_l$; therefore $\tilde{p}(\boldsymbol{x}_l)$ is a Gaussian pdf, and $q(\boldsymbol{x}_l) = \mathrm{CN}(\boldsymbol{x}_l; \widehat{\boldsymbol{x}}_l, \widehat{\boldsymbol{S}}_l^x)$. We stress that the constraint $q(\boldsymbol{x}_l) = \tilde{p}(\boldsymbol{x}_l)$ guarantees the monotonicity of the VB-SAGE algorithm, as we show in the Appendix A. Similarly, we select $\mathcal{Q}(w_l)$ as the set of Gaussian pdfs, i.e., $q(w_l) = \mathrm{CN}(w_l; \widehat{w}_l, \widehat{\Phi}_l)$; notice that $q(w_l) = \tilde{p}(w_l)$ only when $p(w_l | \alpha_l)$ is a Gaussian pdf. For the sparsity parameters $\alpha_l$ we select $\mathcal{Q}(\alpha_l)$ as the set of gamma pdfs, i.e., $q(\alpha_l) = \mathrm{Ga}(\alpha_l; \widehat{a}_l, \widehat{b}_l)$. This choice is dictated by the Gamma distribution being the conjugate prior for the inverse variance of the normal distribution; as a result, in the Gaussian prior case $q(\alpha_l) = \tilde{p}(\alpha_l)$. We select $\mathcal{Q}(\boldsymbol{\theta}_l)$ as the set of Dirac measures on the range of $\boldsymbol{\theta}_l$; thus, $q(\boldsymbol{\theta}_l) = \delta(\boldsymbol{\theta}_l - \widehat{\boldsymbol{\theta}}_l)$. By doing so we restrict ourselves to point estimates of the dispersion parameters.[7] The parameters $\widehat{\boldsymbol{x}}_l, \widehat{\boldsymbol{S}}_l^x, \widehat{\boldsymbol{\theta}}_l, \widehat{a}_l, \widehat{b}_l, \widehat{w}_l$, and $\widehat{\Phi}_l$ are called variational parameters. Obviously, knowing the pdf $q(\boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{x}_l)$ translates into knowing the variational parameters of its factors and vice-versa.

[6]For a given Bayesian network with $\mathcal{O}$ variables, a Markov Blanket of a variable $a$ is the smallest subset of variables $\mathcal{MB}(a) \subseteq \mathcal{O}$ that "shields" $a$ from the rest of the variables $\mathcal{R} = \mathcal{O} \setminus \{a, \mathcal{MB}(a)\}$ in the sense that $p(a | \mathcal{MB}(a), \mathcal{R}) = p(a | \mathcal{MB}(a))$.

[7]Considering more complex forms of $q(\boldsymbol{\theta}_l)$ would require the expectation of $s(\boldsymbol{\theta}_l)$ with respect to $\boldsymbol{\theta}_l$ to be evaluated in the closed form. A detailed study of this case is outside the scope of this paper.

### B. Variational Estimation Expressions

Just like SAGE, the VB-SAGE algorithm is implemented in a sequential manner. For the model with $L$ signal components the algorithm sets $l = 1$ and updates the proxy factors $q(\boldsymbol{x}_1)$, $q(\boldsymbol{\theta}_1)$, $q(w_1)$, and $q(\alpha_1)$ related to the first component, i.e., updates the corresponding variational parameters, based on the currently available estimates of the factors, i.e., the variational parameters, of all $L - 1$ other components. In the same fashion the variational parameters of the component $l = 2$ are updated, and so on, until all $L$ components are considered. The procedure of updating all parameters of all $L$ components in this way constitutes a single update cycle of the algorithm. The update cycles are repeated anew until convergence.

In what follows, we consider the update expressions for the variational parameters $\{\widehat{\boldsymbol{x}}_l, \widehat{\boldsymbol{S}}_l^x, \widehat{\boldsymbol{\theta}}_l, \widehat{a}_l, \widehat{b}_l, \widehat{w}_l, \widehat{\Phi}_l\}$ of the $l$th component only. The updated value of a parameter will be denoted by $(\cdot)'$; let us point out that after $q(\boldsymbol{x}_l)$ has been updated, the other factors related to the component $l$ can be updated in any order.

*1) Estimation of $q(\boldsymbol{x}_l)$:* From the graph in Fig. 1(b), we conclude that $\mathcal{MB}(\boldsymbol{x}_l) = \{\boldsymbol{z}, \boldsymbol{\theta}_{\mathcal{I}}, w_{\mathcal{I}}\}$. Evaluating (7) in this case leads to $\tilde{p}(\boldsymbol{x}_l) \propto p(\boldsymbol{z} | \boldsymbol{x}_l, \widehat{\boldsymbol{\theta}}_{\overline{\mathcal{I}(l)}}, \widehat{w}_{\overline{\mathcal{I}(l)}}) p(\boldsymbol{x}_l | \widehat{w}_l, \widehat{\boldsymbol{\theta}}_l)$. Since the right-hand side is a product of Gaussian pdfs, $\tilde{p}(\boldsymbol{x}_l)$ is as well a Gaussian pdf with the mean and covariance matrix given by

$$\begin{aligned} \widehat{\boldsymbol{x}}_l' &= (1 - \beta_l) \widehat{w}_l \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l) + \beta_l \left( \boldsymbol{z} - \sum_{k \in \overline{\mathcal{I}(l)}} \widehat{w}_k \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_k) \right) \\ \left(\widehat{\boldsymbol{S}}_l^x\right)' &= \left(\boldsymbol{\Delta}_l^{-1} + \boldsymbol{\Sigma}_l^{-1}\right)^{-1}. \end{aligned} \tag{9}$$

Thus, $q(\boldsymbol{x}_l) = \tilde{p}(\boldsymbol{x}_l) = \mathrm{CN}(\boldsymbol{x}_l; \widehat{\boldsymbol{x}}_l, \widehat{\boldsymbol{S}}_l^x)$. The result (9) generalizes that obtained in [3] by accounting for the covariance matrix of $\boldsymbol{x}_l$ and the noise covariance matrix $\boldsymbol{\Sigma}$. Note, however, that the expression for the mean $\boldsymbol{x}_l$ in (9) is identical to that obtained in the SAGE algorithm.

Let us consider the limiting case as $\beta_l \to 1$. It has been shown that for models linear in their parameters the choice $\beta_l = 1$ leads to a fast convergence of the algorithm already in the early iteration steps [28]. This is equivalent to assuming that $\boldsymbol{x}_l = \boldsymbol{z} - \sum_{k \in \overline{\mathcal{I}(l)}} w_k \boldsymbol{s}(\boldsymbol{\theta}_k)$, which was also used as an admissible hidden data in [3]. In this case $\left(\widehat{\boldsymbol{S}}_l^x\right)' \to 0$, so that $q(\boldsymbol{x}_l)$ collapses to a Dirac distribution and $\boldsymbol{\Sigma}_l \to \boldsymbol{\Sigma}$.

*2) Estimation of $q(\boldsymbol{\theta}_l)$:* The Markov blanket of $\boldsymbol{\theta}_l$ is $\mathcal{MB}(\boldsymbol{\theta}_l) = \{w_l, \boldsymbol{x}_l\}$. Here the estimation algorithm profits from the usage of the admissible hidden data $q(\boldsymbol{x}_l)$. Since $q(\boldsymbol{\theta}_l) = \delta(\boldsymbol{\theta}_l - \widehat{\boldsymbol{\theta}}_l)$, finding $q(\boldsymbol{\theta}_l)$ reduces to the computation of $\widehat{\boldsymbol{\theta}}_l$ that maximizes $\tilde{p}(\boldsymbol{\theta}_l)$ given by (7). By noting that $p(\boldsymbol{\theta}_l | w_l, \boldsymbol{x}_l) \propto p(\boldsymbol{x}_l | \boldsymbol{\theta}_l, w_l) p(\boldsymbol{\theta}_l)$ we obtain

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_l' = \underset{\boldsymbol{\theta}_l}{\arg\max} \Big\{ &\log p(\boldsymbol{\theta}_l) + \log p(\widehat{\boldsymbol{x}}_l | \boldsymbol{\theta}_l, \widehat{w}_l) \\ &- \widehat{\Phi}_l \boldsymbol{s}(\boldsymbol{\theta}_l)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}(\boldsymbol{\theta}_l) \Big\}. \end{aligned} \tag{10}$$

Notice that due to $q(w_l)$ being a Gaussian pdf within the VB-SAGE framework, (10) includes a Tikhonov-like regularization term $\widehat{\Phi}_l \boldsymbol{s}(\boldsymbol{\theta}_l)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}(\boldsymbol{\theta}_l)$ with the posterior variance $\widehat{\Phi}_l$ of $w_l$ acting as a regularization constant. Unfortunately, since

$\boldsymbol{s}(\boldsymbol{\theta}_l)$ depends nonlinearly on $\boldsymbol{\theta}_l$, (10) has to be optimized numerically, e.g., using successive line searches where each element of $\widehat{\boldsymbol{\theta}}_l$ is determined separately or using a joint search in which all elements of $\widehat{\boldsymbol{\theta}}_l$ are computed jointly; if derivatives of the objective function (10) with respect to $\boldsymbol{\theta}_l$ are available, gradient-based optimization schemes can also be used.

Typically $q(\boldsymbol{\theta}_l)$ is selected to factorize according to $q(\boldsymbol{\theta}_l) = q(\theta_{l1}) \cdot \ldots \cdot q(\theta_{lR})$, where $R$ is the number of dispersion parameters describing a multipath component.[8] Estimating $\theta_{lr}, r \in \{1, \ldots, R\}$ can be done by evaluating (7) using $\mathcal{MB}(\theta_{lr}) = \{\mathcal{MB}(\boldsymbol{\theta}_l) \bigcup \{\boldsymbol{\theta}_l \backslash \theta_{lr}\}$ and performing a simple line search of the resulting objective function. Notice that the same assumption underpins the SAGE-based estimation of $\boldsymbol{\theta}_l$. The VB-SAGE estimation expression for $\boldsymbol{\theta}_l$ in (10) coincides with that of the standard SAGE when $p(\boldsymbol{\theta}_l)$ is selected non-informative and $q(w_l) = \delta(w_l - \widehat{w}_l)$.

*3) Estimation of $q(w_l)$:* The Markov blanket for $w_l$ is $\mathcal{MB}(w_l) = \{\boldsymbol{\theta}_l, \boldsymbol{x}_l, \alpha_l\}$. Evaluating (7) leads to $\tilde{p}(w_l) \propto p(\widehat{\boldsymbol{x}}_l | w_l, \widehat{\boldsymbol{\theta}}_l) p(w_l | \widehat{\alpha}_l)$. For a given choice of $p(w_l|\alpha_l)$ the moments of $q(w_l) = \mathrm{CN}(w_l; \widehat{w}_l, \widehat{\Phi}_l)$ can be either found in closed form or efficiently approximated. We defer the estimation of these moments to Section IV, where different priors $p(w_l|\alpha_l)$ are discussed.

*4) Estimation of $q(\alpha_l)$:* Here $\mathcal{MB}(\alpha_l) = \{w_l\}$. Observe that in contrast to $q(\boldsymbol{\theta}_l)$ and $q(w_l)$, the admissible hidden data $\boldsymbol{x}_l$ is not in $\mathcal{MB}(\alpha_l)$. This is the result of the Markov chain $\alpha_l \rightarrow w_l \rightarrow \boldsymbol{x}_l$; in fact, $w_l$ is the admissible hidden data for estimating $\alpha_l$ since $p(\boldsymbol{x}_l, w_l | \alpha_l, \boldsymbol{\theta}_l) = p(\boldsymbol{x}_l | w_l, \boldsymbol{\theta}_l) p(w_l | \alpha_l)$ due to the factorization (4). By noting that $p(\alpha_l | w_l) \propto p(w_l | \alpha_l) p(\alpha_l)$, (7) can be rewritten as $\tilde{p}(\alpha_l) \propto p(\alpha_l) \exp(\mathbb{E}_{q(w_l)}\{\log p(w_l|\alpha_l)\})$. Due to the fact that $q(\alpha_l) = \mathrm{Ga}(\alpha_l; \widehat{a}_l, \widehat{b}_l)$, the variational parameters $\widehat{a}_l$ and $\widehat{b}_l$ are found by equating the moments of $q(\alpha_l)$ and $\tilde{p}(\alpha_l)$. Observe that it is the estimation of $q(\alpha_l)$ that eventually leads to the sparse VB-SAGE algorithm. Also notice that the sparsity prior $p(w_l|\alpha_l)$ is a key to the estimation of the sparsity parameters. In the following section, we will consider several choices of $p(w_l|\alpha_l)$ and analyze their effect on sparsity-based model order selection.

## IV. Sparsity Priors for Model Order Selection

In this section we consider three choices for the sparsity prior $p(w_l|\alpha_l)$: i) a Gaussian prior, which leads to the $\ell_2$-type of log-likelihood penalty; ii) a flat prior, obtained as a limiting case of the Gaussian prior when $\alpha_l \rightarrow 0$; and iii) a Laplace prior, which results in the $\ell_1$-type of log-likelihood penalty.

### A. Gaussian Sparsity Prior

The Gaussian sparsity prior is obtained by selecting $p(w_l|\alpha_l) = \mathrm{CN}(w_l; 0, \alpha_l^{-1})$. With this choice it is straightforward to show that $q(w_l) = \tilde{p}(w_l)$ and that

$$\begin{aligned} \widehat{\Phi}_l' &= \left( \widehat{\alpha}_l + \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l) \right)^{-1}, \\ \widehat{w}_l' &= \widehat{\Phi}_l' \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l. \end{aligned} \tag{11}$$

[8]If some of the dispersion parameters are statistically dependent, a structured mean field approximation can be used to account for this dependency by means of an appropriate factorization of the proxy pdf $q(\boldsymbol{\theta}_l)$.

Observe that (11) is merely a regularized least-squares estimate of $\widehat{w}_l'$ given $\widehat{\boldsymbol{x}}_l$ and $\widehat{\boldsymbol{\theta}}_l$ with the regularization parameter $\widehat{\alpha}_l = \mathbb{E}_{q(\alpha_l)}\{\alpha_l\} = \widehat{a}_l \widehat{b}_l$.

The variational parameters $\widehat{a}_l$ and $\widehat{b}_l$ of $q(\alpha_l)$ are found from $\tilde{p}(\alpha)$. This requires the expectation of $|w_l|^2$ to be computed. Doing so leads to the following update expressions:

$$\widehat{a}_l' = a_l + 1, \quad \widehat{b}_l' = b_l + \left( |\widehat{w}_l|^2 + \widehat{\Phi}_l \right). \tag{12}$$

Let us now analyze (12) in more details for the case $a_l = b_l = 0$, i.e., when $p(\alpha_l)$ is non-informative. In this case the mean of $q(\alpha_l)$ is given as

$$\widehat{\alpha}_l' = \frac{1}{|\widehat{w}_l|^2 + \widehat{\Phi}_l}. \tag{13}$$

Note that this result coincides with the EM-based evidence estimation proposed in [11] and [14]. However, in our case both $\widehat{w}_l$ and $\widehat{\Phi}_l$ are estimated using the admissible hidden data $\boldsymbol{x}_l$, as opposed to [11] and [14] where the incomplete data $\boldsymbol{z}$ is used to obtain these estimates. The updating steps in (11) and (13) can be alternatively repeated, while keeping $\widehat{\boldsymbol{x}}_l$ and $\widehat{\boldsymbol{\theta}}_l$ fixed to generate a sequence $\{\widehat{\alpha}_l^{[m]}\}_{m>0}$, where $\widehat{\alpha}_l^{[1]} = \widehat{\alpha}_l'$, $\widehat{\alpha}_l^{[2]} = \widehat{\alpha}_l''$, etc. Note that this updating process makes sense since neither $\boldsymbol{x}_l$ nor $\boldsymbol{\theta}_l$ are in $\mathcal{MB}(\alpha_l)$.[9] Therefore, the corresponding sequence of pdfs $\{q^{[m]}(\alpha_l) = \mathrm{Ga}(\alpha_l; \widehat{a}_l^{[m]}, \widehat{b}_l^{[m]})\}_{m>0}$ necessarily monotonically decreases the variational free energy. Let $\widehat{\alpha}_l^{[\infty]}$ be the stationary point of the sequence $\{\widehat{\alpha}_l^{[m]}\}_{m>0}$ when $m \rightarrow \infty$. In order to simplify the notation we define $\widehat{\boldsymbol{s}}_l \equiv \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)$. By substituting (11) into (13) and solving for $\widehat{\alpha}_l^{[\infty]}$ we obtain (see also [19])

$$\widehat{\alpha}_l^{[\infty]} = \frac{\left( \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l \right)^2}{\left| \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l \right|^2 - \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l}. \tag{14}$$

By definition $\widehat{\alpha}_l^{[\infty]} > 0$, which is satisfied if, and only if,

$$\left| \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l \right|^2 > \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l. \tag{15}$$

By interpreting (13) as a nonlinear dynamic mapping, which at the iteration $m$ maps $\widehat{\alpha}_l^{[m]}$ into $\widehat{\alpha}_l^{[m+1]}$, it can be shown [19] that for $|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l|^2 \leq \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l$ the fixed point of the mapping is at infinity, i.e., $\widehat{\alpha}_l^{[\infty]} = \infty$. As a result, the $l$th signal component can be removed from the model.[10] A similar result was reported in [17] using a non-variational analysis of the marginal log-likelihood function. This allows us to implement model order selection during a parameter update iteration, (i.e., joint multipath component detection and parameter estimation, while still minimizing the variational free energy.

Now, let us reinspect (15). This inequality might at first glance seem a bit counter-intuitive—the quadratic quantity on the right-hand side is compared to the fourth-power quantity on the left-

[9]Notice that this property allows for a straightforward extension of the subsequent analysis to the estimation of sparse virtual channel models [16] since it remains valid even when the dispersion parameters $\boldsymbol{\theta}_l$ are constrained to some resolution grid.

[10]Strictly speaking, this is true only in the case of a non-informative hyperprior $p(\alpha_l)$.

hand side (LHS). In order to better understand the meaning of it, let us divide both sides of (15) by $\left(\widehat{\alpha}_l^{[\infty]}+\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l\right)^2$. It follows that (15) is equivalent to

$$\left|\widehat{w}_l^{[\infty]}\right|^2 > \gamma_l \frac{1}{\widehat{\alpha}_l^{[\infty]}+\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l} = \gamma_l \widehat{\Phi}_l^{[\infty]} \tag{16}$$

where $\gamma_l = \frac{\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l}{\widehat{\alpha}_l^{[\infty]}+\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l} \leq 1$. The LHS term in (16) is an estimate of the posterior variance of $w_l$ scaled by $\gamma_l$. This result leads directly to several important observations:

1) The sparsity parameter $\widehat{\alpha}_l^{[\infty]}$ of the signal component $l$ with $|\widehat{w}_l^{[\infty]}|^2$ smaller than its posterior variance $\widehat{\Phi}_l^{[\infty]}$ scaled by $\gamma_l$ is infinite. Thus, such components can be removed from the model.
2) By multiplying both sides of (16) with $\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l$, we find that this inequality is equivalent to $\widehat{\mathrm{SNR}}_l > \gamma_l^2$, where $\widehat{\mathrm{SNR}}_l = |\widehat{w}_l^{[\infty]}|^2\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l$ is the estimated signal-to-noise ratio (SNR) of the $l$th component. Thus, (15) [and (16)] corresponds to keeping this component provided $\widehat{\mathrm{SNR}}_l > \gamma_l^2$.
3) Condition (15) can be tuned to retain the component provided its estimated SNR is above some predefined level $\mathrm{SNR}' \geq \gamma_l^2$ using the modified condition

$$\left|\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{x}}_l\right|^2 > \widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l \times \frac{\mathrm{SNR}'}{\gamma_l^2}. \tag{17}$$

These results provide us with the required instruments to determine whether a component $l$ with the sparsity parameter $\alpha_l$ should be updated or pruned: if the component $l$ fails to satisfy (15), it is removed since for $\widehat{\alpha}_l^{[\infty]} \to \infty$, $\widehat{w}_l^{[\infty]} \to 0$. In case of (17) we remove the component if its estimated SNR is below some level $\mathrm{SNR}' \geq \gamma_l^2$. Notice that the obtained results allow for an interpretation of the sparsity parameter $\alpha_l$ in terms of estimated SNR of the $l$th component. Thus, model order selection (sparsification) can be realized using simple SNR-guided decisions. It should be stressed that the analysis of (14) is possible only due to the use of the admissible hidden data $\boldsymbol{x}_l$. A standard approach with Gaussian priors [11], [14], [17] requires an $L \times L$ posterior covariance matrix $\widehat{\boldsymbol{\Phi}}$ of the gain coefficient vector $\boldsymbol{w} = [w_1, \ldots, w_L]^T$ to be computed. This significantly complicates the analytical computation of the fixed point $\widehat{\alpha}_l^{[\infty]}$ and its analysis. The sparse VB-SAGE algorithm with Gaussian sparsity prior and model order selection scheme that utilizes (15) or (17) we denote as the VB-SAGE-G algorithm.

### B. Flat Sparsity Prior

In the case where $p(w_l|\alpha_l)$ is chosen to be non-informative, we can still make use of the Bayesian sparsity to estimate the model order. This can be done by using the VB-SAGE-G algorithm in the limiting case as $\widehat{\alpha}_l \to 0$ (i.e., $\widehat{b}_l \to \infty$). Due to the structure of the graph [see Fig. 1(b)], this will only affect the moments of $q(w_l)$, which remain identical to (11) with $\widehat{\alpha}_l = 0$. Clearly, in this case $\gamma_l = 1$ and (16) corresponds to the sparsification of the $l$th component provided $\widehat{\mathrm{SNR}}_l > 1$, i.e., we keep the component when its SNR is above 0 dB. The sparse VB-SAGE algorithm with such model order selection scheme we denote as the VB-SAGE-F algorithm. Observe that (17) can also be used in the case of the VB-SAGE-F algorithm.

### C. Laplace Sparsity Prior (Soft Thresholding)

As the last choice we consider a Laplace prior $p(w_l|\alpha_l)$. We will use an analogous Laplace prior in the complex domain defined as

$$p(w_l|\alpha_l) = \frac{2\alpha_l^2}{\pi}\exp\left(-2\alpha_l|w_l|\right). \tag{18}$$

The mean of $q(w_l)$ can be obtained in closed form:

$$\widehat{w}_l' = \mathrm{sign}\left(\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{x}}_l\right) \times \frac{\max\left(0, \left|\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{x}}_l\right| - \widehat{\alpha}_l\right)}{\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H\boldsymbol{\Sigma}_l^{-1}\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)}. \tag{19}$$

Here $\mathrm{sign}(\cdot)$ is the sign function defined as $\mathrm{sign}(x) = \frac{x}{|x|}$. Expression (19) is also known as a *soft thresholding* rule. To our best knowledge no closed form expression for the posterior variance exists. However, we can approximate it with the result obtained for the real-valued $w_l \neq 0$, which is given as

$$\widehat{\Phi}_l' \approx \left(\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)^H\boldsymbol{\Sigma}_l^{-1}\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)\right)^{-1}. \tag{20}$$

Now we turn to the estimation of the sparsity parameter $\alpha_l$. By plugging (18) in the expression for $\tilde{p}(\alpha_l)$, and ignoring terms independent of $\alpha_l$, we obtain $\tilde{p}(\alpha_l) \propto p(\alpha_l)\alpha_l^2\exp(-2\alpha_l\mathbb{E}_{q(w)}\{|w_l|\})$. Since $q(w_l)$ is Gaussian, $|w_l|$ follows a Rice distribution characterized by the parameters $|\widehat{w}_l|$ (19) and $\sqrt{\frac{\widehat{\Phi}_l}{2}}$ (20). The expectation $\mathbb{E}_{q(w)}\{|w_l|\}$ is then given as $\sqrt{\frac{\widehat{\Phi}_l\pi}{4}}\mathcal{L}_{\frac{1}{2}}\left(\frac{-|\widehat{w}_l|^2}{\widehat{\Phi}_l}\right)$, where $\mathcal{L}_\nu(x)$ denotes the Laguerre polynomial with degree $\nu$. To simplify the estimation of $q(\alpha_l)$, we consider an approximation of $\mathbb{E}_{q(w)}\{|w_l|\}$ as $\frac{|\widehat{w}_l|^2}{\widehat{\Phi}_l} \to \infty$. This approximation is equivalent to assuming a high precision estimate of $w_l$. In this case $\mathbb{E}_{q(w)}\{|w_l|\} = |\widehat{w}_l|$. Then, it is straightforward to show that

$$\widehat{a}_l' = a_l + 2, \quad \widehat{b}_l' = b_l + 2|\widehat{w}_l|. \tag{21}$$

By selecting a non-informative prior $p(\alpha_l)$, the update expression for the mean $\widehat{\alpha}_l' = \frac{\widehat{a}_l'}{\widehat{b}_l'}$ simplifies to

$$\widehat{\alpha}_l' = 1/|\widehat{w}_l|. \tag{22}$$

Similar to the Gaussian prior case we analyze the fixed point $\widehat{\alpha}_l^{[\infty]}$ of (22). We define $\widehat{\boldsymbol{s}}_l \equiv \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)$ to simplify the notation. Combining (22) and (19) leads to

$$\widehat{\alpha}_l^{[\infty]} = \frac{\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{s}}_l}{\max\left(0, \left|\widehat{\boldsymbol{s}}_l^H\boldsymbol{\Sigma}_l^{-1}\widehat{\boldsymbol{x}}_l\right| - \alpha_l^{[\infty]}\right)}. \tag{23}$$

Assuming that $|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l| > \alpha_l^{[\infty]}$ (otherwise $\alpha_l^{[\infty]} = \infty$), we solve for $\alpha_l^{[\infty]}$. Doing so yields two solutions:

$$\alpha_{l,+}^{[\infty]} = \frac{1}{2}\left(\left|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l\right| + \mu_l\right) \tag{24}$$

$$\alpha_{l,-}^{[\infty]} = \frac{1}{2}\left(\left|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l\right| - \mu_l\right) \tag{25}$$

where $\mu_l = \sqrt{\left|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{x}_l\right|^2 - 4\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l}$. Furthermore, we see that a necessary and sufficient condition for the fixed points to be real is that

$$\left|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l\right|^2 \geq 4\, \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l. \tag{26}$$

Components that do not satisfy (26) are removed. Note that both fixed points are feasible. We have always empirically observed that when the initial $q(\alpha_l)$ is chosen such that $\widehat{\alpha}_l = 0$, iterations (22) either diverge ($\alpha_l^{[\infty]} = \infty$) or converge to the closest (smallest) feasible solution given by (25). The properties of the second stationary point are subject to further investigations left outside the scope of this paper. The sparse VB-SAGE algorithm with Laplace sparsity priors that makes use of (26) for model order selection we denote as the VB-SAGE-L algorithm.
Similarly to (16) it can be shown that (26) is equivalent to

$$\left|\widehat{w}_l^{[\infty]}\right|^2 \geq 4\gamma_l^2 \frac{1}{\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l} = 4\gamma_l^2 \widehat{\Phi}_l^{[\infty]} \tag{27}$$

with $\gamma_l = \frac{|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l| - \widehat{\alpha}_l^{[\infty]}}{|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l|} \leq 1$. In the same way, (26) and (27) are equivalent to keeping the component provided $\widehat{\mathrm{SNR}}_l \geq 4\gamma_l^2$, where $\widehat{\mathrm{SNR}}_l = \left|\widehat{w}_l^{[\infty]}\right|^2 \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l$ is the estimated component SNR. Note that (26) and (27) are the Laplace-prior equivalent conditions of (15) and (16) respectively for the Gaussian prior. Although the pruning conditions are formally similar, they differ in their numerical values: the moments of $q(w_l)$ are estimated differently computed in these two schemes; as a result, the estimates of the admissible hidden data $\boldsymbol{x}_l$ for the VB-SAGE-L and VB-SAGE-G algorithms are also different; in addition, the scaling factor $\gamma_l$ in (27) is computed differently from that in (16). It should also be mentioned that as $\alpha_l \to 0$ the VB-SAGE-L algorithm converges to the VB-SAGE-F algorithm.

Similarly to (17), (26) can be tuned to keep the component when its estimated SNR is above some predefined level $\mathrm{SNR}' \geq 4\gamma_l^2$ using the modified condition

$$\left|\widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l\right|^2 \geq \widehat{\boldsymbol{s}}_l^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{s}}_l \times \frac{\mathrm{SNR}'}{\gamma_l^2}. \tag{28}$$

## V. Implementation and Initialization of the Algorithm

### A. Summary of the Algorithm

Let us now summarize the main steps of the proposed algorithm. For the moment we assume that at some iteration $j$ the proxy factors $q(\boldsymbol{x}_l)$, $q(\boldsymbol{\theta}_l)$, $q(w_l)$, and $q(\alpha_l)$, $l \in \{1, \ldots, \widehat{L}\}$, are known for the $\widehat{L}$ components. A single update iteration for the component $l$ is summarized in Algorithm 1.

**Algorithm 1:** Update iteration for the component $l$

Update $q(\boldsymbol{x}_l)$ from (9)

Update $q(\boldsymbol{\theta}_l)$ from (10) and evaluate $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l')$

**if** Condition (17) or (28) are TRUE **then**

Update $q(\alpha_l)$ from (14) (VB-SAGE-G) or (25) (VB-SAGE-L)

Update $q(w_l)$ from (11) (VB-SAGE-G, -F) or (19) (VB-SAGE-L)

$\widehat{L}' \leftarrow \widehat{L}$

**else**

Remove the $l$th component; $\widehat{L}' \leftarrow \widehat{L} - 1$

**end if**

This update iteration is repeated for all components in a round-robin fashion, which constitutes a single update cycle of the algorithm. The update cycles are then repeated until the number of components and their variational parameters converge. Observe that the number of components might be reduced during one update cycle: at each iteration the updated multipath component undergoes a test specified by (17) or (28). When the corresponding condition is not satisfied the component is removed. The model order might also be increased by adding new components. Details of this procedure are outlined in Section V-D.

### B. Algorithm Initialization

We propose a simple bottom-up initialization strategy, which allows us to infer the initial variational parameters from the observation $\boldsymbol{z}$ by starting with an empty model, i.e., assuming all variational parameters to be 0. The first component is initialized by letting $\widehat{\boldsymbol{x}}_1' = \boldsymbol{z}$ and applying the initialization loop shown in Algorithm 2. Observe that using $\widehat{\boldsymbol{x}}_l'$ the dispersion parameters $\widehat{\boldsymbol{\theta}}_l'$ are initialized using a simple beamformer and the obtained estimate of $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l')$ is plugged in (15) (in the Gaussian prior case) or in (26) (in the Laplace prior case) to determine whether the initialized component should be kept in the model. When the test fails, the initialization stops. It should be stressed that the use of (15) or (26) during the initialization is optional and may be omitted if an overcomplete channel representation is desired. The components with large sparsity parameters will then be pruned later during the update iterations. This initialization strategy is similar to the successive interference cancellation scheme proposed in [3] and [5]. The number of initialization iterations (i.e., the initial number of signal components) can be either fixed to $L_{\max}$, or inferred automatically by repeating the initialization iterations until the pruning condition (15) [or (26)] fails at some

iteration.[11] In our implementation of the algorithm, we use a combination of the two methods by limiting the maximum number of initial components to $L_{\max}$. The application of the VB-SAGE algorithm requires the specification of several free parameters. Specifically, one has to select the covariance matrix of the additive noise and the parameter $\beta_l$ in the definition of the admissible hidden data. The choice of these parameters is described below.

**Algorithm 2:** Algorithm initialization

Set $l \leftarrow 1$; initialize $q(\boldsymbol{x}_l)$: $\widehat{\boldsymbol{x}}_1' \leftarrow \boldsymbol{z}$

**while** Continue initialization **do**

Initialize $q(\boldsymbol{\theta}_l)$ by computing $\widehat{\boldsymbol{\theta}}_l' = \arg\max_{\boldsymbol{\theta}_l} \frac{|\boldsymbol{s}(\boldsymbol{\theta}_l)^H \boldsymbol{\Sigma}_l^{-1} \widehat{\boldsymbol{x}}_l'|}{\boldsymbol{s}(\boldsymbol{\theta}_l)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}(\boldsymbol{\theta}_l)}$

**if** Condition (15) (VB-SAGE-G, -F) or (26) (VB-SAGE-L) are TRUE **then**

Initialize $q(w)$ from (11) with $\widehat{\alpha}_l = 0$

Initialize $q(\alpha_l)$ from (12) (VB-SAGE-G) or (21) (VB-SAGE-L)

$\widehat{L}' = l$; $l \leftarrow l + 1$

$\widehat{\boldsymbol{x}}_l' \leftarrow \boldsymbol{z} - \sum_{k=1}^{l-1} \widehat{w}_k' \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_k')$;

**else**

Stop initialization: $\widehat{L}' = l - 1$

**end if**

**end while**

*1) Noise Statistics:* A crucial part of the initialization procedure is the accurate estimation of the variance of the additive noise $\boldsymbol{\xi}$. Logically, when the noise level is high, we tend to put less "trust" in the estimates of the signal parameters and thus sparsify components more aggressively.

In many cases estimates of the noise variance can be derived from the signal itself. Specifically, the noise variance can be estimated from the tail of the measured CIR. Alternatively, the noise variance can be estimated from the residual signal obtained after completion of the initialization step. In our work we use the former initialization strategy.

*2) Selecting $\boldsymbol{\Sigma}_l$:* The obtained sparsity expressions for model order selection all depend on the covariance matrix of the additive noise $\boldsymbol{\xi}_l$ associated with the $l$th multipath component. The covariance matrix $\boldsymbol{\Sigma}_l$ is related to the total covariance matrix $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_l = \beta_l \boldsymbol{\Sigma}$, where $\beta_l$ is the noise splitting parameter introduced in the definition of the admissible hidden data (3). In the SAGE algorithm applied to the estimation of superimposed signal parameters [3] this parameter was set to $\beta_l = 1$; we also adopt this choice. Obviously, in this case $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}$ and $\widehat{\mathrm{SNR}}_l = |w_l|^2 \boldsymbol{s}(\boldsymbol{\theta}_l)^H \boldsymbol{\Sigma}^{-1} \boldsymbol{s}(\boldsymbol{\theta}_l)$.

[11]We suggest to use (15) or (26) instead of their modified versions (17) and (28), since this allows for the inclusion of even the weakest components during the initialization.

### C. Stopping Criterion for the Update Cycles

The iterative nature of the algorithm requires a stopping criterion for the variational parameter updates. In our implementation we use the following simple criterion: the estimation iterations are terminated when: i) the number of signal components stabilizes; and ii) the maximum change of the components in $\{\boldsymbol{\Omega}, \boldsymbol{\alpha}\}$ between two consecutive update cycles is less than $0.01\%$.

### D. Adaptive Model Order Estimation

The structure of the estimation algorithm also allows for increasing the model order. Increasing the model order might be useful when $L_{\max}$ is selected too small so that not all physical multipath components might have been discovered. Alternatively, new components might also appear in time-varying scenarios. The new components can be initialized from the residual signal. After the model fitting has been performed at some update cycle, e.g., $j$, the residual $\widehat{\boldsymbol{x}}_{L+1}' = \boldsymbol{z} - \sum_{l=1}^{L} \widehat{w}_l \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)$ is computed and used to initialize new components as explained in Section V-B. Essentially, the residual signal can be used at any stage of the algorithm to initialize new components.

### E. Estimation Uncertainty and Selection of the Sensitivity Level $\mathrm{SNR}'$

There are four main sources of uncertainty in model-based multipath estimation: i) the inaccuracy of the specular model (1) in representing reality (e.g., in the presence of diffuse components); ii) the error in calibrating the measurement equipment, which results in an error in the specification of the mapping $u(t) \mapsto \boldsymbol{s}(t, \boldsymbol{\theta}_l)$; iii) the discrete-time approximation (2) of the model; and iv) the discrete optimization that is typically necessary due to the nonlinearity of the model versus some of its parameters. All these aspects have a significant impact on the model order estimation. Any deviation from the "true" model [effects i) and ii)] and inaccuracies in the parameter estimates $\widehat{\boldsymbol{\Omega}}$ [due to iii) and iv)] result in a residual error, manifesting itself as a contribution from fictive additional components. If no penalization of the parameter log-likelihood is used, this error leads to additional signal components being detected, especially in high SNR regime. These non-physical components are numerical artifacts; they do not correspond to any real multipath components. Moreover, these fictive components, which are typically much weaker than the real specular components, create pseudo-clusters since typically their parameters are highly correlated. In the case of the VB-SAGE-G, VB-SAGE-F and VB-SAGE-L algorithms, the artifacts can be efficiently controlled using the pruning conditions (17) and (28) with an appropriately chosen sensitivity level $\mathrm{SNR}'$. The sensitivity level can be set globally, or can be tuned individually to each multipath component. We propose the following implementation of individual tuning.

First, we consider the impact of all aforementioned inaccuracies together. This approach is motivated by experimental evidence indicating that: i) each type of inaccuracies has a non-negligible effect on channel estimation and ii) that these effects are difficult to quantify and also to separate. Second, we assume

that—due to these inaccuracies—the residual error contributed by a given estimated multipath component is proportional to the sample of the delay power profile at the component delay. Indeed, it makes sense to presume that the stronger a multipath component is, the larger the residual error due to calibration and discretization error is. This rationale leads us to select $\mathrm{SNR}' \equiv \mathrm{SNR}'(\tau)$ proportional to a low-pass filtered version of the delay power profile $\mathrm{DPP}(\tau)$. In Section VI-B we discuss how this scheme is applied to measured CIRs.

Note that there are also alternative approaches to account for the inaccuracy of the specular model. In [32] the authors propose a method that jointly estimates the specular multipath components and the diffuse component, called dense multipath component (DMC), in a time-variant MIMO context. The parameters of the components (direction of departure (DoD), direction of arrival (DoA), relative delay, Doppler frequency, polarimetric path gain) are estimated using an extended Kalman filter built around a dynamic model of these parameters. The parameters of the DMC are computed from the residual signal resulting after subtracting the estimated specular components from the observed signal. Obviously, an accurate estimation of the specular part of the channel plays a vital role here. We now discuss the main differences between of the sparse VB-SAGE algorithm proposed here and the method published in [32]. First, both algorithms apply a path pruning algorithm that relies on comparing the path weight to a threshold. The pruning algorithm proposed here is based on a Bayesian sparsity framework, while that used in [32] implements the Wald test. This leads to different ways of computing the pruning threshold and the signals compared to this threshold. Second, the sparse VB-SAGE algorithm does not make any particular assumption on the structure of the DMC. Experimental evidence suggests that the DoD-DoA-delay power spectrum characterizing the DMC typically does not factorize, up to a proportionality constant, in the product of the corresponding DoD, DoA, and delay spectra, as implied by the Kronecker factorization of the transmit-array–receive-array–frequency covariance matrix assumed in [32]. The inherent directionality of the radio channel, which holds for both specular components and diffuse components, translates in power spots scattered in the DoD-DoA-delay space that cannot be represented by the above factored spectrum [see also Fig. 5(d)–(f) and Fig. 6(d)–(f)]. This observation, combined with the other early mentioned model inaccuracies, has motivated the empirical method based on the selected $\mathrm{SNR}'(\tau)$ threshold. Finally, the sparse VB-SAGE is derived and applied in a time-invariant SIMO scenario with only one polarization considered. As mentioned earlier it can be easily extended to time-variant MIMO scenario including full path polarization, provided the propagation constellation is stationary. Extension to the time-variant scenario with changing propagation constellation as considered in [32] will require further work. A thorough investigation is needed to assess the pros and cons of the model order selection methods applied in the channel estimation proposed in [32] and in the sparse VB-SAGE algorithm. This study is, however, beyond the scope of this paper.

## VI. Application of the Sparse VB-SAGE Algorithm to the Estimation of Wireless Channels

### A. Synthetic Channel Responses

We first demonstrate the performance of the algorithm with synthetic channel responses generated according to model (2). We use a sounding sequence with $M = 63$ chips and a square-root-raised-cosine shaping pulse $p(t)$ with a duration $T_\mathrm{p} = 10$ nsec and a roll-off factor 0.25. A horizontal-only propagation scenario is considered with a received replica of the transmitted signal represented as $w_l \boldsymbol{s}(t, \boldsymbol{\theta}_l) = w_l \boldsymbol{c}(\phi_l) u(t - \tau_l)$ where $w_l$, $\phi_l$, and $\tau_l$ denote respectively the complex gain, the azimuthal direction and the relative delay of the $l$th multipath component. Thus, $\boldsymbol{\theta}_l = \{\phi_l, \tau_l\}$. The $M_\mathrm{r}$-dimensional complex vector $\boldsymbol{c}(\phi_l) = [c_1(\phi_l), \ldots, c_{M_\mathrm{r}}(\phi_l)]^T$ is the steering vector of the array [3]. We assume a linear array with $M_\mathrm{r} = 16$ ideal isotropic sensors spaced half a wavelength apart. The parameters of the multipath components are chosen by randomly drawing samples from the corresponding distributions: delays $\tau_l$ and angles $\phi_l, l = 1, \ldots, L$, are drawn uniformly in the interval [0.03, 0.255] $\mu$s and $[\frac{-\pi}{2}, \frac{\pi}{2}]$, respectively. For generating the multipath gains $w_l$ we follow two scenarios. In the first scenario we generate the gains as $w_l = \sqrt{P}\mathrm{e}^{j\eta_l}$, where $P$ is some positive constant and $\eta_l, l = 1, \ldots, L$, are independent random phases uniformly distributed in the interval [0, $2\pi$). This ensures that all multipath components have the same power $P$ and therefore the same per-component SNR. In the second scenario the values of $w_l, l = 1, \ldots, L$, are independently drawn from a complex Gaussian distribution with the pdf $\mathrm{CN}(w_l; 0, P'\mathrm{e}^{\frac{-\tau_l}{\tau_\mathrm{s}}})$, where $P'$ is some positive constant and $\tau_\mathrm{s}$ is the delay spread set to $\frac{T_\mathrm{b}}{4}$. In this case the distribution of the component gains $w_l$ is conditioned on the delay $\tau_l$ such that the average received power decays exponentially as the delay increases. The later choice approximates better the real behavior of component powers versus delay. At the same time it demonstrates the performance of the algorithm under conditions with changing per-component SNR.

By sampling $\boldsymbol{z}(t)$ with a sampling period $T_\mathrm{s}$ we obtain the equivalent discrete-time formulation (2) with $N$ samples per channel. The samples of the received signal are recorded over the time window $T_\mathrm{b} = 0.63$ $\mu$s (i.e., $N_\mathrm{u} = 1$) at a rate $\frac{1}{T_\mathrm{s}} = 200$ MHz. In the simulations we set the number of specular components to $L = 20$. By fixing $L$ we aim to demonstrate the possible bias of the model order selection mechanism. Additive noise $\boldsymbol{\xi}$ is assumed to be white with covariance matrix $\boldsymbol{\Sigma} = \sigma_\xi^2 \boldsymbol{I}$. Different SNR conditions are simulated. The considered SNR is the averaged per-component SNR defined as

$$\mathrm{SNR} = \frac{1}{L} \sum_l \frac{|w_l|^2 \, \|\boldsymbol{s}(\boldsymbol{\theta}_l)\|^2}{\sigma_\xi^2}.$$

With this setting the estimation step (10) is implemented as a sequence of two numerical optimizations. For instance, the estimation of $\tau_l$ with $\mathcal{MB}(\tau_l) = \{\boldsymbol{x}_l, w_l, \phi_l\}$ is performed first as

$$\tau_l' = \arg\max_{\tau_l} \Big\{ \log p\left(\widehat{\boldsymbol{x}}_l' | \tau_l, \widehat{\phi}_l, \widehat{w}_l\right) - \widehat{\Phi}_l \boldsymbol{s}(\tau_l, \widehat{\phi}_l)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}(\tau_l, \widehat{\phi}_l) + \log p(\tau_l) \Big\} \quad (29)$$

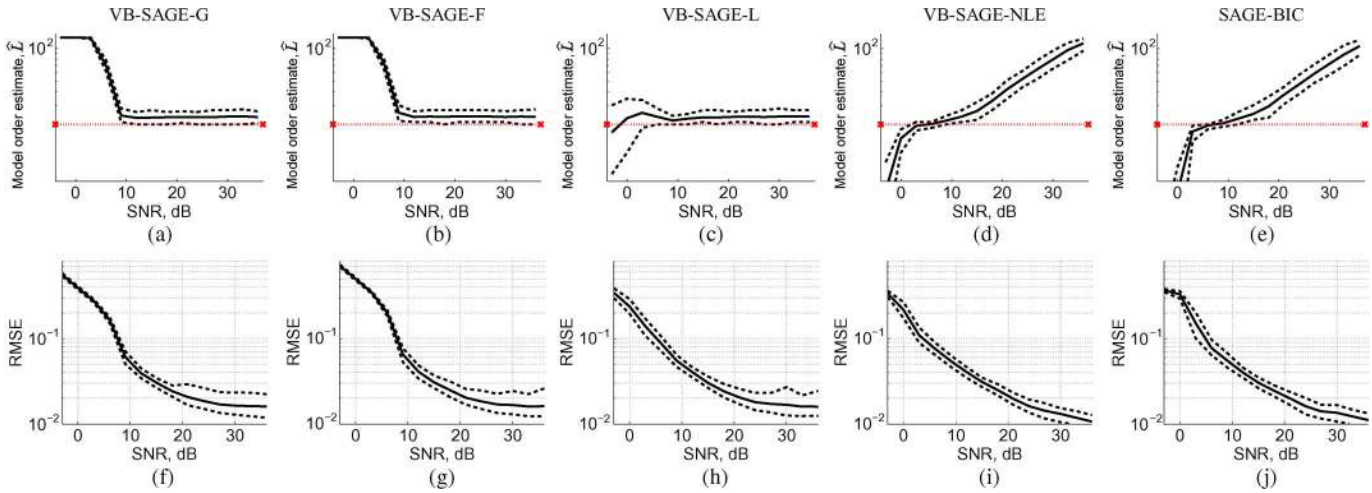


Fig. 2. Performance of the proposed estimation algorithms applied to synthetic channels with equal component power. Estimation of model order $\widehat{L}$ (a)–(e), and the achieved RMSE between the synthetic and reconstructed responses (f)–(j). The true number of components is $L = 20$ (dotted line in upper plots). The solid lines denote the averaged estimates of the corresponding parameters. Upper and lower dotted lines denote the 5th and 95th percentiles of the estimates, respectively.

followed by the estimation of the azimuth $\phi_l$ with $\mathcal{MB}(\phi_l) = \{\boldsymbol{x}_l, w_l, \tau_l\}$ as

$$\widehat{\phi}_l' = \arg\max_{\phi_l} \Big\{ \log p\left(\widehat{\boldsymbol{x}}_l' | \widehat{\tau}_l', \phi_l, \widehat{w}_l\right) - \widehat{\Phi}_l \boldsymbol{s}\left(\widehat{\tau}_l', \phi_l\right)^H \boldsymbol{\Sigma}_l^{-1} \boldsymbol{s}\left(\widehat{\tau}_l', \phi_l\right) + \log p(\phi_l) \Big\}. \quad (30)$$

Optimizations (29) and (30) are performed using a simple line search on a grid followed by polynomial interpolation to improve the precision of the estimates. For the initialization of the algorithm we use the scheme described in Section V-B. The maximum number of initialized components is set to $L_{\max} = N$. We use the modified pruning conditions (17) for the VB-SAGE-G and VB-SAGE-F schemes and (28) for the VB-SAGE-L algorithm with $\mathrm{SNR}'$ set to the true SNR used in the simulations. This setting demonstrates the performance of the algorithms when the true per-component SNR is known. In particular, it allows us to investigate how the modified pruning conditions can be used to control the estimation artifacts.

We compare five estimation algorithms: i) VB-SAGE-G; ii) VB-SAGE-F; iii) VB-SAGE-L; iv) the SAGE algorithm [3] with Bayesian information criterion for model order selection (SAGE-BIC); and v) the VB-SAGE algorithm with the negative log-evidence (NLE) approach for model order selection (VB-SAGE-NLE) [19]. The NLE is equivalent to the Bayesian interpretation of the normalized ML model order selection [7], [9]. For SAGE-BIC and VB-SAGE-NLE, we set the initial number of components to the number of samples $N$.

We first consider the simulation scenario where all components have the same power. The corresponding results, averaged over 200 Monte Carlo runs, are summarized in Fig. 2. It can be seen that VB-SAGE-G, VB-SAGE-F, and VB-SAGE-L clearly outperform the other two methods, with VB-SAGE-L exhibiting the best performance. Notice that (17) in VB-SAGE-G and VB-SAGE-F fails for low SNR; also the initial number of components (126 in this case) remains unchanged during the update iterations. The VB-SAGE-L algorithm, however, does not exhibit such behavior. Nonetheless, all three methods have a small positive model order bias in the high SNR regime.

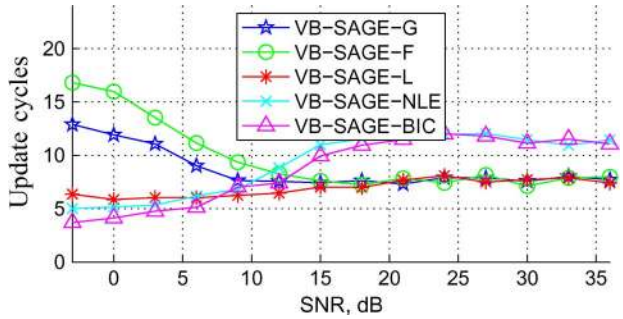


Fig. 3. Averaged number of update cycles versus the averaged per-component SNR.

VB-SAGE-NLE and SAGE-BIC perform reasonably only in the limited SNR range 8–14 dB and fail as the SNR increases beyond. The reason for this is an inadequate penalization of the parameter likelihood, which leads to the introduction of estimation artifacts. Specifically, the selected sampling rates of the processed signals limit the precision in the estimation of the dispersion parameters $\boldsymbol{\Theta}$ of the multipath components. As a result the mean-squared error of these estimates exhibits a floor at high SNR. These estimates are obtained by optimizing parameter-specific objective functions, cf. (29) and (30), which in a real implementation are computed from discrete signals. As a consequence, the objective functions need to be interpolated between their computed samples in these optimization procedures. It is the error resulting from these interpolations that leads to the flooring of the estimate errors at high SNR regime. The residual errors of the dispersion parameters translate into residual interference that may manifest itself as fictive components if not handled appropriately. This effect can also be seen as a basis mismatch problem that leads to an overestimation of true model sparsity [26]. The use of adjusted pruning conditions in case of VB-SAGE-G, -F, and -L algorithms allows for a better control over the estimation artifacts. This, however, leads to a floor of the RMSE between the synthetic and reconstructed channel responses at high SNR, as seen in Figs. 2(f), (g), and (h). In contrast, VB-SAGE-NLE and VB-SAGE-BIC do not exhibit this behavior of RMSE, albeit

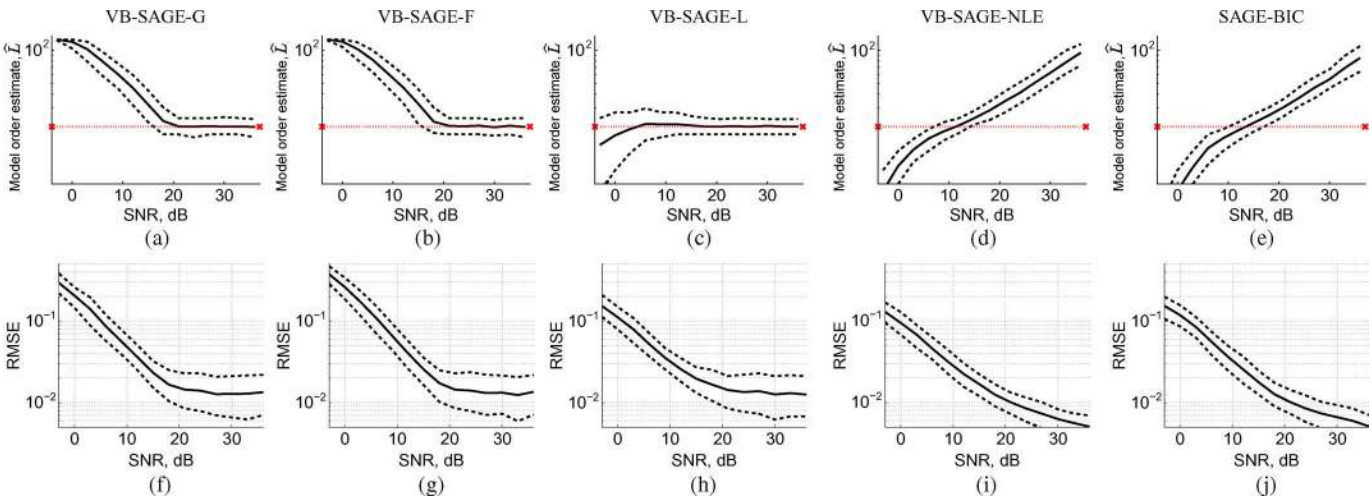


Fig. 4. Performance of the proposed estimation algorithms applied to synthetic channels with exponentially decaying component power. Estimation of model order $\widehat{L}$ (a)–(e), and the achieved RMSE between the synthetic and reconstructed responses (f)–(j). The true number of components is $L = 20$ (dotted line in upper plots). The solid lines denote the averaged estimates of the corresponding parameters. Upper and lower dotted lines denote the 5th and 95th percentiles of the estimates, respectively.

at the expense of introducing more and more fictive multipath components to compensate for multipath parameter estimation errors as the SNR increases.[12] Increasing the number of samples $N$ while keeping $T_{\mathrm{b}}$ fixed and increasing the number of antenna elements reduces the noise RMSE floor since the multipath dispersion parameters can be estimated with greater precision.

Obviously, the model order estimate has a significant impact on the convergence speed of the algorithm. Fig. 3 depicts the averaged number of update cycles versus SNR for the five investigated channel estimation schemes. We see here that for an SNR above 12 dB the VB-SAGE-G, -F, and -L schemes outperform the other estimation schemes, with the convergence rate of the VB-SAGE-L algorithm being almost independent of the SNR. Notice that the overestimation of the model order with VB-SAGE-NLE and SAGE-BIC leads to a significant increase of the number of iterations as the SNR increases.

Let us now consider the second scenario where the component power decreases exponentially versus delay. The results are reported in Fig. 4. A picture similar to that of the equal-power case is observed here. The performance of VB-SAGE-L is clearly better than that of the other tested schemes. In this setting both VB-SAGE-G and VB-SAGE-F require higher SNR to bring the estimated model order within the range of the true number of components. Notice that the VB-SAGE-G, -F, and -L methods are no longer biased and on average estimate the correct number of components.

### *B. Estimation of Measured Wireless Channels*

We now investigate the performance of the VB-SAGE-L algorithm applied to the estimation of measured wireless channel responses collected in an indoor environment. The measurements were done with the MIMO channel sounder PropSound manufactured by Elektrobit Oy. Details on the measurement campaign can be found in [34]. To compute the results presented in this paper we used a portion of the measurement data that corresponds to a line-of-sight scenario. The sounder operated at the center frequency 5.25 Ghz with a chip period $T_{\mathrm{p}} = 10$ ns. We used the 9 dual-polarized elements of the bottom ring of the receive antenna array and all 25 dual-polarized elements of the transmit array (see Fig. 1c in [34]), i.e., $M_{\mathrm{r}} = 18$ and $M_{\mathrm{t}} = 50$. The sounding sequence consisted of $M = 255$ chips, resulting in a burst waveform of duration $T_{\mathrm{b}} = 2.55\,\mu$s. One burst waveform was sent to sound each channel corresponding to a pair or transmit antenna and receive antenna. The received signal was sampled with the period $T_{\mathrm{s}} = \frac{T_{\mathrm{p}}}{2}$ (i.e., 2 samples/chip).

The estimation results obtained using the VB-SAGE-L algorithm are compared to Bartlett estimates [33]. We report only the azimuthal information of the estimated multipath components. In order to minimize the effect of estimation artifacts we make use of (28). The sensitivity level $\mathrm{SNR}'$ is computed from the estimated delay power profile as described in Section V-E: a smoothed estimate of the delay power profile $\mathrm{DPP}(\tau)$ is normalized with the estimated additive noise variance $\sigma_\xi^2$; the sensitivity $\mathrm{SNR}'(\tau)$ is then defined as[13] $\mathrm{SNR}'(\tau) = \frac{\frac{\mathrm{DPP}(\tau)}{\sigma_\xi^2}}{10^{1.5}}$. This setting allows for the detection (removal) of components at a certain delay with power above (below) a threshold set 15 dB below the received power at that delay. The algorithm is initialized as described in Section V-B. To initialize $\tau_l$'s we partition the DPP in 8 delay segments covering the delay interval [10,360] ns. Then, using (29) and (30) we initialize at most 7 components per segment,[14] which results in $L_{\max} = 56$. For the used sensitivity level $\mathrm{SNR}'(\tau)$ the algorithm estimates $L = 18$ components. The parameter estimates of these components are summarized in Figs. 5 and 6.

[12]Note, however, that the same effect is observed with VB-SAGE-G and VB-SAGE-L when $\mathrm{SNR}'$ is not used to enforce sparsity and correct for model order estimation errors.

[13]A possible extension, not considered here due to space limitations, would consists in making $\mathrm{SNR}'$ both delay and direction dependent.

[14]The initialization of the multipath components located in a delay segment is interrupted when the pruning condition (26) fails.

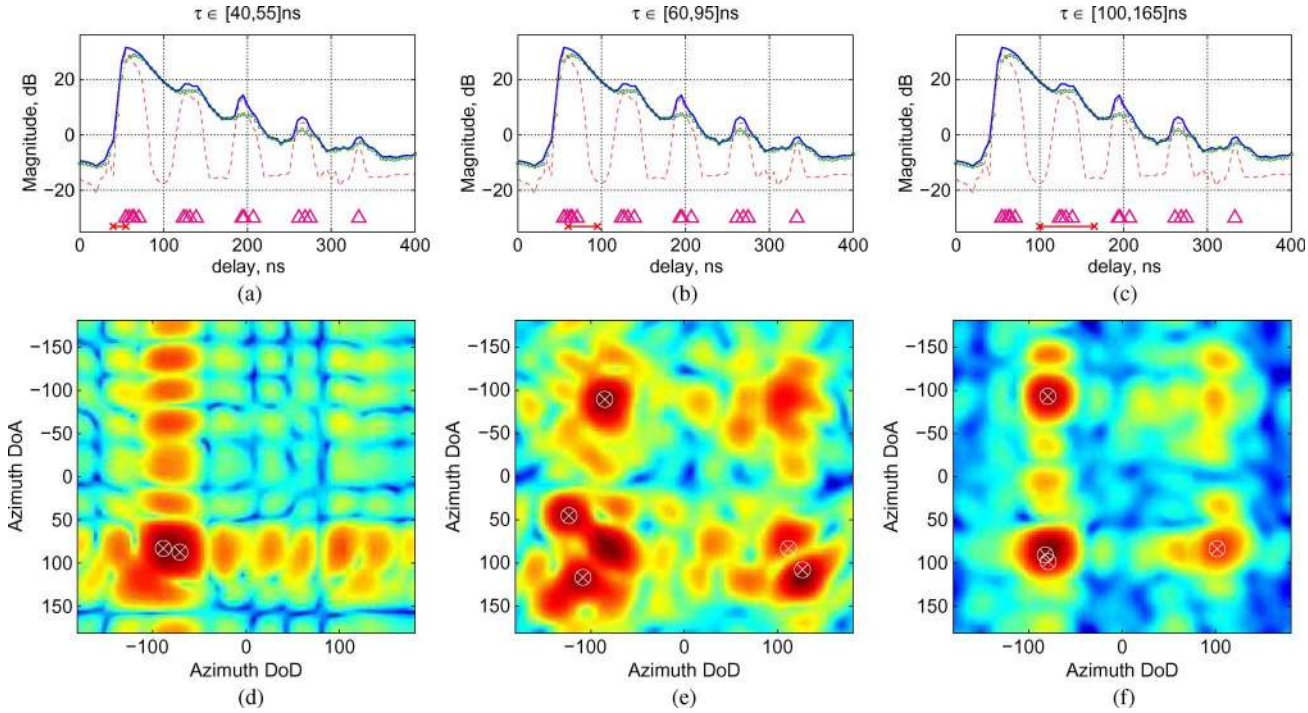


Fig. 5. (a)–(c) Bartlett estimates (solid line) and model-based estimates (dashed line) of the delay power profile; dotted lines denote the estimated delay power profile of the residual $\boldsymbol{\xi}$; triangles denote the delays of the estimated components; (d)–(f) normalized Bartlett estimates of the azimuth of arrival (DoA) and departure (DoD) for the selected delay intervals (denoted by crosses in figures (a)–(c), respectively; crossed circles denote the azimuths of the estimated components.

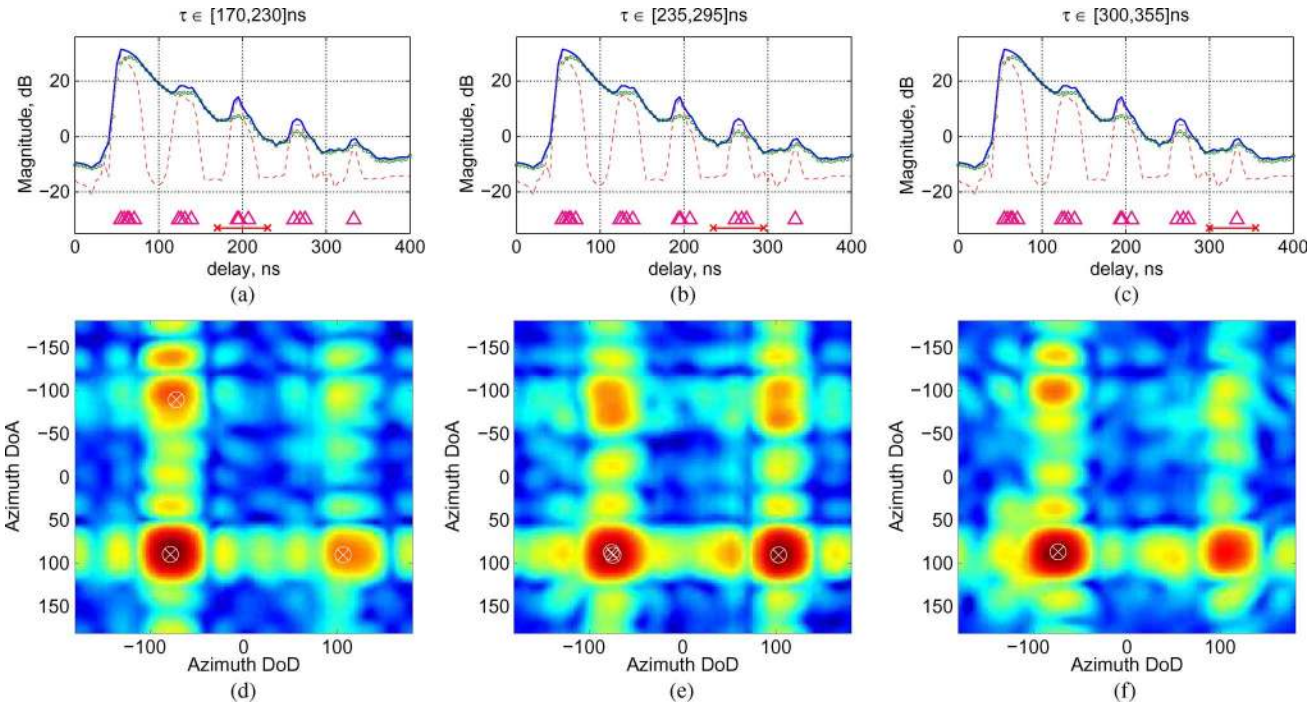


Fig. 6. (a)–(c) Bartlett estimates (solid line) and model-based estimates (dashed line) of the delay power profile; dotted lines denote the estimated delay power profile of the residual $\boldsymbol{\xi}$; triangles denote the delays of the estimated components; (d)–(f) normalized Bartlett estimates of the azimuth of arrival (DoA) and departure (DoD) for the selected delay intervals (denoted by crosses in figures (a)–(c), respectively; crossed circles denote the azimuths of the estimated components.

Investigations, not reported due to space limitation, show that the estimated multipath components can be associated to propagation paths computed from the geometry of the environment using ray-tracing. Due to the delay-dependent sensitivity level $\mathrm{SNR}'(\tau)$ very weak components in the tail of the delay response are also detected. Their positions coincide well with the maxima of the Bartlett spectra. We also note that not all "footprints" in the Bartlett spectra have been identified as multipaths. This is due to the component magnitudes being below the detection sensitivity of the algorithm; also, some of the footprints observed in the Bartlett spectra are likely due to side lobes caused by the system response and thus may not correspond to any true physical multipath component.

## VII. Conclusion

This contribution proposes a new algorithm that estimates the number of relevant multipath components in the response of radio channels and the parameters of these components within the Bayesian framework. High-resolution estimation of the multipath components is performed using the VB-SAGE algorithm—a new extension of the traditional SAGE algorithm—which allows for computing estimates of the posterior pdfs of the component parameters, rather than parameter point estimates. By introducing sparsity priors for the multipath component gains, the sparse VB-SAGE algorithm estimates the posterior pdfs of the component parameters jointly with the posterior pdfs of the sparsity parameters by minimizing the variational free energy. The pdfs of the parameters of a single component are updated at each iteration of the algorithm, with the iterations cycling through the components. Due to the monotonicity property of the VB-SAGE algorithm, the free energy is non-decreasing versus the iterations.

Several sparsity priors are considered: Gaussian, flat, and Laplace priors. The admissible hidden data introduced in the VB-SAGE algorithm lead to simple and easy to interpret component pruning rules/conditions for these priors. Theses conditions are shown to be equivalent to removing signal components based on comparison of the per-component SNR with a given threshold. This threshold can be set for all components or tailored for each component individually.

The sparse VB-SAGE algorithm is applied to the estimation of the multipath components in the response of synthetic and measured wireless multipath channels. We show by means of Monte Carlo simulations that the sparsity-based model order selection methods with sensitivity-adjusted pruning conditions outperform the Bayesian Information Criterion and the negative log-evidence model order selection criterion. The latter approaches fail since, due to various effects (calibration errors, finite precision in the discretization process, diffuse scattering, etc.) leading to a model mismatch, numerical artifacts are introduced, which lead to a decreasing RMSE at the expense of an increased model order. In case of estimation of wireless channels this is highly undesirable, since the estimated artifacts have no physical meaning. The proposed modifications of the pruning conditions allow for correcting for possible model order estimation bias due to modeling mismatch. Making use of the Laplace prior results in the best performance among the tested methods. Simulations show that for low SNR the VB-SAGE algorithm with Laplace sparsity priors, which we refer to as the VB-SAGE-L algorithm, keeps only reliably estimated components, while successfully removing the artifacts. The VB-SAGE-L algorithm also exhibits the fastest convergence as compared to the other tested algorithms with the same stopping criterion.

We apply the VB-SAGE-L algorithm to the estimation of the multipath components in measured channel impulse responses. In order to minimize the effects of model mismatch, the detector sensitivity $\mathrm{SNR}'$ is adjusted based on an estimate of the delay power profile. Since the artifacts are typically more pronounced in delay ranges associated with high received power, a smoothed version of the delay power profile can be used as an indicator of the received power versus propagation delay. Investigations, not reported in this paper due to space limitation, show that the estimated multipath components can be associated to propagation paths computed from the geometry of the environment using ray-tracing.

The sparse VB-SAGE algorithm provides a new and effective tool for efficient estimation of wireless channels. Its flexibility and its iterative structure make it very attractive for many applications in wireless communications: analysis and estimation of complex MIMO channel configurations in channel sounding and MIMO radars, channel estimation in iterative receivers performing joint channel estimation and data decoding, as well as extraction of location-dependent features of the radio channel for localization purposes.

## Appendix A
## Monotonicity Property of the VB-SAGE Algorithm

In what follows, we assume that the variational approximating pdf (6) and its factors are selected as outlined in Section III-A and $\beta_l$ is set to 1.

Define $\mathcal{A}_l = \{w_l, \boldsymbol{\theta}_l, \alpha_l\}$ as the set of parameters associated with the $l$th multipath component and $\mathcal{R}_l = \{w_k, \boldsymbol{\theta}_k, \alpha_k; k \in \overline{\mathcal{I}(l)}\}$ as the set of the other multipath parameters. We assume that $q(\mathcal{A}_l, \mathcal{R}_l) = q(\mathcal{A}_l)q(\mathcal{R}_l)$. It is straightforward to show that minimizing the free energy $\mathcal{F}(q(\mathcal{A}_l, \mathcal{R}_l)||p(\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l))$ with respect to $q(\mathcal{A}_l)$ is equivalent to minimizing $\mathcal{F}(q(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l))$ with $\tilde{p}(\boldsymbol{z}, \mathcal{A}_l) \propto \exp(\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\})$. The VB-SAGE algorithm facilitates this optimization using the admissible hidden data $\boldsymbol{x}_l$ in (3). Consider the equality $p(\boldsymbol{x}_l, \boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l) = p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)p(\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)$. By combining this equality with the factorization (4) and computing the expectation with respect to $\boldsymbol{x}_l$ and $\mathcal{R}_l$ we obtain

$$\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\} = \mathbb{E}_{q(\boldsymbol{x}_l)}\{\log p(\boldsymbol{x}_l, \mathcal{A}_l)\} - \mathbb{E}_{q(\boldsymbol{x}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\} + \mathrm{const.}$$

where const is a term independent of $\mathcal{A}_l$. Define now $\tilde{\tilde{p}}(\mathcal{A}_l) \propto \exp(\mathbb{E}_{q(\boldsymbol{x}_l)}\{\log p(\boldsymbol{x}_l, \mathcal{A}_l)\})$. Observe that $p(\boldsymbol{x}_l, \mathcal{A}_l)$ is a function of the admissible hidden data and the $l$th multipath component parameters. Now, the free energy with respect to $\mathcal{A}_l$ can be rewritten as

$$\mathcal{F}\left(q(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l)\right) = \mathcal{F}\left(q(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l)\right) - \mathbb{E}_{q(\boldsymbol{x}_l)}\mathbb{E}_{q(\mathcal{A}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\} + \mathrm{const.} \quad (31)$$

Minimizing $\mathcal{F}(q(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l))$ is typically simpler as compared to minimizing $\mathcal{F}(q(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l))$. However, whether $\mathcal{F}(q(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l))$ decreases as $\mathcal{F}(q(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l))$ decreases ultimately depends on the term $\mathbb{E}_{q(\boldsymbol{x}_l)}\mathbb{E}_{q(\mathcal{A}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\}$ in (31).

Let $q(\mathcal{A}_l)$ denote an existing (old) estimate of $\mathcal{A}_l$, and let $q'(\mathcal{A}_l)$ be the new minimizer of $\mathcal{F}(q(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l))$. A current estimate $q(\boldsymbol{x}_l)$ of the admissible hidden data posterior pdf is given by (7), i.e., $q(\boldsymbol{x}_l) = \tilde{p}(\boldsymbol{x}_l) \propto \exp(\mathbb{E}_{q(\mathcal{A}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\})$, since $\mathcal{MB}(\boldsymbol{x}_l) = \{\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l\}$. Note that it is easy to show that $\log\tilde{p}(\boldsymbol{x}_l)$ must be quadratic in $\boldsymbol{x}_l$. Similarly we define $\tilde{p}'(\boldsymbol{x}_l) \propto \exp\{\mathrm{E}_{q'(\mathcal{A}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l)\}\}$. With these settings it follows that

$$\begin{aligned}\mathcal{F}\left(q(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l)\right) - \mathcal{F}\left(q'(\mathcal{A}_l)||\tilde{p}(\boldsymbol{z}, \mathcal{A}_l)\right) \\ = \mathcal{F}\left(q(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l)\right) - \mathcal{F}\left(q'(\mathcal{A}_l)||\tilde{\tilde{p}}(\mathcal{A}_l)\right) \\ + \mathrm{D}_{\mathrm{KL}}\left(\tilde{p}(\boldsymbol{x}_l)||\tilde{p}'(\boldsymbol{x}_l)\right) \geq 0. \quad (32)\end{aligned}$$

Result (32) expresses the monotonicity property of the VB-SAGE algorithm. Furthermore, $q(\boldsymbol{x}_l) = \tilde{p}(\boldsymbol{x}_l) \propto \exp\{\mathbb{E}_{q(\mathcal{A}_l)}\mathbb{E}_{q(\mathcal{R}_l)}\{\log p(\boldsymbol{x}_l|\boldsymbol{z}, \mathcal{A}_l, \mathcal{R}_l\}\}$ is a sufficient condition that guarantees the monotonicity of the VB-SAGE algorithm for our estimation problem.

## REFERENCES


[1] T. S. Rappaport, *Wireless Communications. Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall PTR, 2002.

[2] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, pp. 67–94, Jul. 1996.

[3] B. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, "Channel parameter estimation in mobile radio environments using the SAGE algorithm," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 434–450, Mar. 1999.

[4] O. Besson and P. Stoica, "Decoupled estimation of DOA and angular spread for a spatially distributed source," *IEEE Trans. Signal Process.*, vol. 48, no. 7, pp. 1872–1882, 2000.

[5] A. Richter, "Estimation of radio channel parameters: Models and algorithms," Ph.D. dissertation, Tech. Univ. Ilmenau, Ilmenau, Germany, 2005.

[6] H. Akaike, "A new look at the statistical model identification," *Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[7] J. I. Myung, D. J. Navarro, and M. A. Pitt, "Model selection by normalized maximum likelihood," *J. Math. Psychol.*, vol. 50, pp. 167–179, 2005.

[8] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[9] A. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model order estimation," *Int. Statist. Rev.*, vol. 69, no. 2, pp. 185–212, 2000.

[10] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[11] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[12] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[13] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.* , vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[14] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, June 2001.

[15] M. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, 2003.

[16] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[17] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proc. 9th Int. Workshop on Artif. Intell. Statist.*, Key West, FL, Jan. 2003.

[18] R. Neal, *Bayesian Learning for Neural Networks*, ser. Lecture Notes in Stat.. New York: Springer-Verlag, 1996, vol. 118.

[19] D. Shutin, G. Kubin, and B. H. Fleury, "Application of the evidence procedure to the analysis of wireless channels," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–23, 2007.

[20] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal Process.*, vol. 86, no. 3, pp. 549–571, 2006.

[21] J. W. Wallace and M. A. Jensen, "Sparse power angle spectrum estimation," *IEEE Trans. Antennas Propag.*, vol. 57, no. 8, pp. 2452–2460, Aug. 2009.

[22] P. Zhao and B. Yu, "On model selection consistency of LASSO," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[23] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1$ minimization," *Proc. Nat. Acad. of Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003.

[24] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Statist. Soc.*, vol. 58, pp. 267–288, 1994.

[25] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.* , vol. 1, no. 4, pp. 586–597, Dec. 2007.

[26] Y. Chi, A. Pezeshki, L. Scharf, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.* , 2010.

[27] M. J. Beal, "Variational algorithm for approximate Bayesian inference," Ph.D. dissertation, Univ. College London, London, U.K., 2003.

[28] J. Fessler and A. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Process.*, vol. 42, pp. 2664–2677, Oct. 1994.

[29] D. Shutin and H. Koeppl, "Application of the evidence procedure to linear problems in signal processing," in *Proc. 24th Int. Workshop on Bayesian Infer. Max. Entr. Methods in Sci. Eng.*, Jul. 2004, pp. 124–127.

[30] D. J. C. MacKay, "Bayesian methods for backpropagation networks," in *Models of Neural Networks III*, E. Domany, J. L. van Hemmen, and K. Schulten, Eds. New York: Springer-Verlag, 1994, ch. 6, pp. 211–254.

[31] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. : Springer, Aug. 2006.

[32] J. Salmi, A. Richter, and V. Koivunen, "Detection and tracking of MIMO propagation path parameters using state-space approach," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1538–1550, Apr. 2009.

[33] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley, 2002.

[34] N. Czink, E. Bonek, L. Hentila, J.-P. Nuutinen, and J. Ylitalo, "Cluster-based MIMO channel model parameters extracted from indoor time-variant measurements," in *Proc. Global Telecommun. Conf. (GLOBECOM '06)*, Nov. 27, 2006.


**Dmitriy Shutin** (S'02–M'06) received the Master's degree in computer science in 2000 from Dniepropetrovsk State University, Ukraine, and the Ph.D. degree in electrical engineering from Graz University of Technology, Graz, Austria, in 2006.

During 2001–2006 and 2006–2009, he was a Teaching Assistant and an Assistant Professor, respectively, with the Signal Processing and Speech Communication Laboratory, Graz University of Technology. Since 2009, he has been a Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ. His current research interests include stochastic modeling and estimation of the radio channel for MIMO applications, distributed inference techniques in agent/sensor networks, radar signal processing, and statistical pattern recognition.

Dr. Shutin was a recipient of the Best Student Paper Award at the 2005 IEEE International Conference on Information, Communications and Signal Processing (ICICS). In 2009 he was awarded the Erwin Schrödinger Research Fellowship.

**Bernard H. Fleury** (M'97–SM'99) received the Diploma in electrical engineering and mathematics in 1978 and 1990, respectively, and the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology Zurich (ETHZ) in 1990.

Since 1997, he has been with the Department of Electronic Systems, Aalborg University, Denmark, as a Professor of communication theory. He is the Head of the Section Navigation and Communications, which is one of the eight laboratories of this Department. From 2006 to 2009, he was a Key Researcher with Telecommunications Research Center Vienna (FTW), Austria. During 1978–1985 and 1992–1996, he was a Teaching Assistant and a Senior Research Associate, respectively, with the Communication Technology Laboratory, ETHZ. Between 1988 and 1992, he was a Research Assistant with the Statistical Seminar at ETHZ. His research interests cover numerous aspects within communication theory and signal processing, mainly for wireless communications. His current activities include stochastic modeling and estimation of the radio channel especially for MIMO applications in fast time-varying environments, iterative message-passing processing with focus on the design of efficient feasible architectures for wireless receivers, localization techniques in wireless terrestrial systems, and radar signal processing. He has authored and coauthored more than 110 publications in these areas. He has developed, with his staff, a high-resolution method for the estimation of radio channel parameters that has found a wide application and has inspired similar estimation techniques both in academia and in industry.