Machine Learning Techniques in Optical Communication

Darko Zibar, *Member, IEEE*, Molly Piels, Rasmus Jones, and Christian G. Schäeffer, *Member, IEEE*

(Invited Paper)

***Abstract*—Machine learning techniques relevant for nonlinearity mitigation, carrier recovery, and nanoscale device characterization are reviewed and employed. Markov Chain Monte Carlo in combination with Bayesian filtering is employed within the nonlinear state-space framework and demonstrated for parameter estimation. It is shown that the time-varying effects of cross-phase modulation (XPM) induced polarization scattering and phase noise can be formulated within the nonlinear state-space model (SSM). This allows for tracking and compensation of the XPM induced impairments by employing approximate stochastic filtering methods such as extended Kalman or particle filtering. The achievable gains are dependent on the autocorrelation (AC) function properties of the impairments under consideration which is strongly dependent on the transmissions scenario. The gain of the compensation method are therefore investigated by varying the parameters of the AC function describing XPM-induced polarization scattering and phase noise. It is shown that an increase in the nonlinear tolerance of more than 2 dB is achievable for 32 Gbaud QPSK and 16-quadratic-amplitude modulation (QAM). It is also reviewed how laser rate equations can be formulated within the nonlinear state-space framework which allows for tracking of nonLorentzian laser phase noise lineshapes. It is experimentally demonstrated for 28 Gbaud 16-QAM signals that if the laser phase noise shape strongly deviates from the Lorentzian, phase noise tracking algorithms employing rate equation-based SSM result in a significant performance improvement (>8 dB) compared to traditional approaches using digital phase-locked loop. Finally, Gaussian mixture model is reviewed and employed for nonlinear phase noise compensation and characterization of nanoscale devices structure variations.**

***Index Terms*—Bayesian filtering, expectation maximization (EM), machine learning, Monte Carlo methods, optical communication.**

I. Introduction

HIGH baud rate (>30 Gbaud) optical communication systems employing modulation formats up to 64-quadratic-amplitude modulation (QAM) are soon to become commercially available [1]. In the research community, there is currently focus on increasing the signal baud rate beyond 100 Gbaud [2]–[4] and also realizing single-carrier Tb/s systems [5]. Even though there has been a lot of progress in the field lately, there are still many challenges remaining to be solved in terms of fibre nonlinearity compensation (the maximum amount of information transmitted over a specific distance is limited by the optical fibre nonlinearity [6]), phase noise compensation (integration of semiconductor laser on a transceiver chip imposes certain restrictions on the achievable laser noise properties [7]) and equalization enhanced phase noise for high-baud rate systems [3].

To be more specific, next generation of signal processing algorithms should be able to cope with: finite memory induced by the interaction between accumulated chromatic dispersion, Kerr nonlinearity and noise [8], [9], system identification of various channel impairments and components [10], phase tracking for non-Lorentzian lineshape optical sources [7] and modulation for nonlinear communication channels [11]– [13]. This requires a departure from the signal processing methods designed for linear channel and exploration of non-linear statistical signal processing methods offered by the machine learning community [14], [15]. In this paper, we present a nonlinear state-space framework in conjunction with sequential Monte Carlo methods that can be used for system identification and tracking of time-varying parameters. The framework allows for the dynamics of the optical channel and components to be included in the formulation of the signal processing algorithms. We demonstrate the benefits of the proposed approach for XPM induced polarization scattering and phase noise, carrier synchronization for non-Lorentzian lineshape optical sources. Additionally, we present Gaussian mixture model (GMM) and its application in optimum symbol detection in the presence of nonlinear phase noise and also characterization of silicon photonics process variations. The paper is an extension of [10] with a significant contribution to the theoretical part and addition of new results with respect to mitigation of XPM induced impairments and density estimations.

The remainder of this paper is organized as follows. In Section II, a method for system identification of nonlinear state-space model (SSM) is presented. This is achieved by employing Metropolis-Hastings (MH) algorithm in conjunction with the state estimation. The pseudo-code for the system identification method is also presented. In Section III, it is shown that the effects of XPM induced polarization scattering and phase noise can be included in the state-space framework by modeling it as an autoregressive (AR) processes. The extended Kalman filter (EKF) is formulated for the joint tracking of cross-phase modulation (XPM) induced polarization scattering and phase noise. The performance of the EKF is compared with the results obtained by particle filtering which is a more general method of dealing with nonlinear SSM. Next, in Section IV, it is reviewed how laser dynamics expressed by rate equations can be

Manuscript received October 11, 2015; revised November 23, 2015 and December 10, 2015; accepted December 10, 2015. Date of publication December 16, 2015; date of current version March 3, 2016.

D. Zibar, M. Piels, and R. Jones are with the DTU Fotonik, Department of Photonics Engineering, Technical University of Denmark, Kgs. Lyngby DK-2800, Denmark (e-mail: dazi@fotonik.dtu.dk; mopi@fotonik.dtu.dk; rasjones1990@gmail.com).

C. G. Schaeffer is with the Faculty of Electrical Engineering, Helmut Schmidt University, Hamburg 22008, Germany (e-mail: cgs@hsu-hh.de).

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/JLT.2015.2508502

included into the SSM and allow for tracking of non-Lorentzian lineshapes. The improvement in the system performance, when using laser rate equation based carrier recovery, is quantified based on the experimental results. Finally, in Section V, basic concepts behind the GMM are reviewed. The application of GMM in combination of expectation maximization (EM) algorithm is demonstrated to compute the optimum decision boundaries for constellations impaired by the nonlinear phase noise. It is also briefly explained how GMM can be applied to characterize probability density function associated with structure variation of photonic integrated circuits. The conclusions are presented in Section VI.

II. Bayesian Filtering and Parameter Estimation

The probabilistic SSM offers a general and very powerful tool to learn, model, and analyze optical communication systems and components. The probabilistic SSM specified at a discrete time instant $k \in \mathbb{N}$ is expressed as:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \tag{1}$$

$$\mathbf{x}_k \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta}) \tag{2}$$

$$\mathbf{y}_k \sim p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta}) \tag{3}$$

where k is the discrete time index. The state $\mathbf{x}_k \in \mathbb{R}^n$, where n is the dimension of the state vector, can represent various dynamical parameters of the system such as: the transmitted data sequence, amplitude noise, phase noise, equalization enhanced phase noise, polarization mode dispersion, XPM induced effects, nonlinear phase noise and carrier density. The variable $\mathbf{y}_k \in \mathbb{R}^m$, represents observable variables which are the samples after the analog-to-digital converter in coherent optical systems, and m is the dimension of the measurement vector. The stochastic dynamics of the state vector $\mathbf{x}_k$ are characterized by $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, which describes the transition probability associated with the uncertainties of the state vector. The measured data is characterized by the conditional probability density function $p(\mathbf{y}_k|\mathbf{x}_k)$ and is determined by the measurements noise (noise generated due to electrical and optical amplification). The probability density functions $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and $p(\mathbf{y}_k|\mathbf{x}_k)$ are parameterized by mean vector, $\boldsymbol{\mu}$, covariance matrix, $\boldsymbol{\Sigma}$, or in some cases hyper-parameters Υ. Those parameters may not be known, and the coefficients of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and Υ are then grouped together into a vector of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the dimension of the vector. The parameters of the vector $\boldsymbol{\theta}$ are assumed to be unobserved random variables with prior probability density function $p(\boldsymbol{\theta})$.

The central idea in Bayesian filtering is to compute the posterior distribution of $\mathbf{x}_{1:K}$ and $\boldsymbol{\theta}$, given the measurements up to time step $k = K$, i.e., $\mathbf{y}_{1:K}$. The posterior distribution can then be either directly or indirectly used to estimate the mean of the state vector and parameters. The posterior distribution is expressed using Baye's rule:

$$p(\mathbf{x}_{1:K}, \boldsymbol{\theta}|\mathbf{y}_{1:K}) = \frac{p(\mathbf{y}_{1:K}|\mathbf{x}_{1:K}, \boldsymbol{\theta})p(\mathbf{x}_{1:K}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}_{1:K})}. \tag{4}$$

It is challenging to compute the joint posterior distribution of the states and the parameters. It is more computationally effective if the problem is separated in computing the posterior of the parameters and states separately. The posterior distribution of the parameters $\boldsymbol{\theta}$ is obtained by integrating out the states:

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:K}) = \int p(\mathbf{x}_{1:k}, \boldsymbol{\theta}|\mathbf{y}_{1:K}) d\mathbf{x}_{1:K} = \frac{p(\mathbf{y}_{1:K}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}_{1:K})}. \tag{5}$$

The expression $p(\mathbf{y}_{1:K})$ does not depend on the parameter vector $\boldsymbol{\theta}$ and it only acts as a normalization constant. Therefore, it can be omitted. The expression for the likelihood of the observed data conditioned on the parameters, $p(\mathbf{y}_{1:K}|\boldsymbol{\theta})$, is not suitable for recursive computational framework and therefore needs to be factorized:

$$p(\mathbf{y}_{1:K}|\boldsymbol{\theta}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}) \tag{6}$$

where $p(\mathbf{y}_k|\mathbf{y}_{:k-1}, \boldsymbol{\theta})$ is computed as [16]:

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}) = \int p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta}) p(\mathbf{x}_k|\mathbf{y}_{k-1}, \boldsymbol{\theta}) d\mathbf{x}_k. \tag{7}$$

The parameter vector $\boldsymbol{\theta}$ is then obtained by maximizing the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}_{1:K})$:

$$\boldsymbol{\theta}_{\text{est}}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\{p(\boldsymbol{\theta}|\mathbf{y}_{1:K})\}. \tag{8}$$

The maximum likelihood (ML) of the parameter vector $\boldsymbol{\theta}$ is obtained by assuming a uniform prior for $p(\boldsymbol{\theta})$ and performing the following minimization:

$$\boldsymbol{\theta}_{\text{est}}^{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmin}}\{-\log p(\mathbf{y}_{1:K}|\boldsymbol{\theta})\}. \tag{9}$$

Equation (7) is central as it connects parameter estimation and state estimation through $p(\mathbf{y}_k|\mathbf{y}_{1:k}, \boldsymbol{\theta})$ and $p(\mathbf{x}_k|\mathbf{y}_{k-1}, \boldsymbol{\theta})$, respectively. This allows for joint parameter and state estimation. The expression for $p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta})$ is given by the SSM, equation (3), and the expression for $p(\mathbf{x}_k|\mathbf{y}_{k-1}, \boldsymbol{\theta})$ is provided by the Kalman or particle filter. For the linear SSM with Gaussian densities, an analytical expression for $p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ is obtained [16]:

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}) = N(\mathbf{y}_k|\overbrace{\mathbf{H}\mathbf{m}_k^p}^{\boldsymbol{\mu}}, \overbrace{\mathbf{H}\mathbf{P}_k^p\mathbf{H}^T + \boldsymbol{\Sigma}_k}^{\boldsymbol{\Sigma}}) \tag{10}$$

where $N(\cdot)$ denotes Gaussian distribution, $\mathbf{m}_k^p$ and $\mathbf{P}_k^p$ are predicted state mean value and covariance, respectively, both available from Kalman filtering equations [15]. The measurement model matrix $\mathbf{H}$ is available directly from the specified SSM (3). $\boldsymbol{\Sigma}_k$ is the measurement noise covariance matrix associated with $p(\mathbf{y}_k|\mathbf{x}_k)$. For a more a general case of a SSM employing particle filtering for state estimation, the expression for $p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ becomes:

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}) = \sum_{i=1}^{N} w_{k-1}^i p(\mathbf{y}_k|\mathbf{x}_k^{(i)}, \boldsymbol{\theta}) \tag{11}$$

where N is the total number of the particles and w_{k-1}^i are weights from the the previous iteration $k-1$. Pseudo code for computing $p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ is specified by Algorithm 1:

Algorithm 1 Calculate $p(\mathbf{y}_{1:K}|\boldsymbol{\theta})$

specify: $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]$
for m = 1:M **do**
  $\theta = \theta_m$
  for k=1:K **do**
    run state estimation via Kalman or particle filter
    compute: (10) or (11)
    compute: $L_m(\theta) = L_m(\theta) - \log p(\mathbf{y}_k|\mathbf{y}_{k-1}, \theta)$
  end for
end for
$L(\theta) = [L_1, \ldots, L_M]$
$p(\mathbf{y}_{1:K}|\theta) = \exp\{-L(\theta)\}$

From the practical point of view, computation of $p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ or the corresponding log-likelihood function $L(\boldsymbol{\theta}) = \sum_{k=1}^{K} -\log p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ is very useful as it provides the knowledge about the actual shape of the function that needs to be maximized or minimized. The algorithm 1 is a brute force approach of estimating parameters as $p(\mathbf{y}_{1:K}|\boldsymbol{\theta})$ is swept for all possible guesses of $\boldsymbol{\theta}$ and then the parameters θ_m that maximize $p(\mathbf{y}_{1:K}|\boldsymbol{\theta})$ are chosen. The method works well, in terms of computational efficiency, if very few parameters are to be estimated. For more complex cases, Monte Carlo Markov Chain (MCMC) methods need to be employed. In the next section, the MCMC method for parameter estimation is reviewed. There are also many other approaches for parameter estimation within the SSM such as prediction error method employing least squares methods, dual Kalman filtering and data augmentation based on the EM algorithm [17].

A. Monte Carlo Markov Chain

In general, MCMC refers to a technique of generating samples, $\mathbf{y}_k$, from a probability distribution, $p(\mathbf{y})$, based on constructing a Markov chain, such that the generated samples, $\mathbf{y}_k$, have the desired distribution as the target distribution $p(\mathbf{y})$. For the considered case in this paper, this implies that we would like to draw samples, $\boldsymbol{\theta}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}_{1:K})$ and subsequently estimate the parameters by computing the mean:

$$\mathrm{E}[\boldsymbol{\theta}|\mathbf{y}_{1:K}] = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}_{1:K}) d\boldsymbol{\theta}. \tag{12}$$

The integration in equation (12) is in most cases of interest intractable, however, if we can draw independent samples, $\boldsymbol{\theta}_m$ from the distribution $p(\boldsymbol{\theta}|\mathbf{y}_{1:K})$, the integration in (12) can be approximated:

$$\mathrm{E}[\boldsymbol{\theta}|\mathbf{y}_{1:K}] \approx \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_m. \tag{13}$$

There is a large number of iterative methods on how to draw samples from a distribution, and an interested reader is referred to [14]. In this paper, we will focus on the MH algorithm. The MH algorithm is the most common type of MCMC algorithm and also most common types of MCMC algorithms can be interpreted as a special case of the MH algorithm. For a tutorial on the MH algorithm see [18]. The main idea behind the MH algorithm is since it is not possible to draw samples directly from the distribution $p(\boldsymbol{\theta}|\mathbf{y}_{1:K})$, a proposal distribution, $q(\boldsymbol{\theta})$, from which samples $\boldsymbol{\theta}_m$ can be more easily drawn is needed. Once the samples, $\boldsymbol{\theta}_m$, have been drawn, $p(\boldsymbol{\theta}_m|\mathbf{y}_{1:K})$ is evaluated and in course terms related to $p(\boldsymbol{\theta}_{m-1}|\mathbf{y}_{1:K})$, i.e., $p(\boldsymbol{\theta}_m|\mathbf{y}_{1:K})/p(\boldsymbol{\theta}_{m-1}|\mathbf{y}_{1:K})$ from the iteration $m-1$. The new sample, $\boldsymbol{\theta}_m$ is accepted with probability expressed by A:

$$A = \min\left\{\frac{p(\mathbf{y}_{1:K}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m)q(\boldsymbol{\theta}_{m-1}|\boldsymbol{\theta}_m)}{p(\mathbf{y}_{1:K}|\boldsymbol{\theta}_{m-1})p(\boldsymbol{\theta}_{m-1})q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{m-1})}, 1\right\}. \tag{14}$$

Algorithm 2 MH for parameter estimation

specify proposal distribution $q(\boldsymbol{\theta})$: $N(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $U(\boldsymbol{\theta}|\mathbf{a}, \mathbf{b})$
specify prior distribution $p(\boldsymbol{\theta})$
for m = 1:M **do**
  for p = 1:P **do**
    sample: $\boldsymbol{\theta}_m \sim q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{m-1})$
    for k = 1:K **do**
      run state estimation via Kalman or particle filter
      compute: (10) or (11)
      compute: $L_m = L_m - \log p(\mathbf{y}_k|\mathbf{y}_{k-1}, \theta)$
    end for
    evaluate $\alpha = \exp[L_m + \log p(\boldsymbol{\theta}_m) + \log q(\boldsymbol{\theta}_{m-1}|\boldsymbol{\theta}_m) - L_{m-1} - \log p(\boldsymbol{\theta}_{m-1}) - \log q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{m-1})]$
    sample: $\mathbf{u} \sim U(\mathbf{u}|0, 1)$
    if $\mathbf{u} < \min\{0, \alpha\}$ **then**
      $\boldsymbol{\theta}_m = \boldsymbol{\theta}_m$
    else
      $\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1}$
    end if
  end for
end for
evaluate autocorrelation, $R(n)$, of generated samples $\boldsymbol{\theta}_m$
evaluate mean: $\mathrm{E}[\boldsymbol{\theta}] = \mathrm{mean}\,\boldsymbol{\theta}(M_{\mathrm{init}} : M)$

Algorithm 2 shows the basic pseudo-code for MH algorithm for joint parameter and state estimation.

In Algorithm 2, M is the total number of samples, (equivalent to a number of iterations), that we would like to draw and use for parameter estimation. The variable P denotes the lag parameter and helps the chain improve the acceptance rate. In any case, $P << M$. In the following, we will explain in more details Algorithm 2. In general, it is quite challenging to find a good proposal distribution as it is problem dependent. However, most commonly used proposal distributions are Gaussian and uniform. If the Gaussian distribution is selected as the proposal distribution, the free parameter, since it is conditioned on $\boldsymbol{\mu} = \boldsymbol{\theta}_m$, is the covariance matrix $\boldsymbol{\Sigma}$ and for the uniform distribution it is the interval $[\mathbf{a}, \mathbf{b}]$. The prior distribution, which is related to the initialization algorithm, is also important. Typically, a prior, distribution, $p(\boldsymbol{\theta})$, is chosen such that the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:K})$ has the same functional form as the prior, i.e., so called conjugate prior distribution [14].

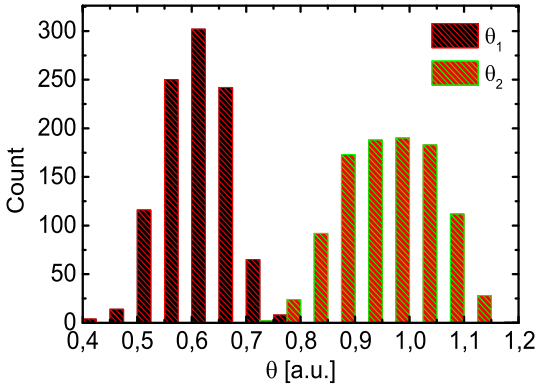

Fig. 1. Histogram of the estimated parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2]$ for $M = 5000$ and $P = 1$.

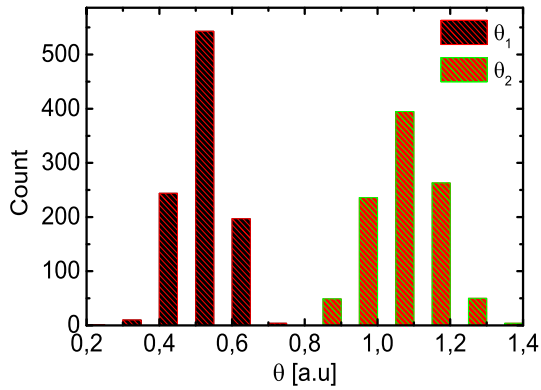

Fig. 2. Histogram of the estimated parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2]$ for $M = 500$ and $P = 10$.

In this paper, it is assumed that $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta})$ and $p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta})$ are Gaussian distribution. In most cases, the unknown parameter vector $\boldsymbol{\theta}$ contains coefficients of the mean and covariance matrix of the distributions describing the state-transition and measurement equations: $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and $p(\mathbf{y}_k|\mathbf{x}_k)$. For the case, when the mean of distributions $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and $p(\mathbf{y}_k|\mathbf{x}_k)$, denoted $\boldsymbol{\mu}^{\text{state}}$ and $\boldsymbol{\mu}^{\text{meas}}$ are unknown, the unknown parameter vector is denoted: $\boldsymbol{\theta} = [\mu_1^{\text{state}}, \ldots \mu_1^{\text{state}}, \ldots \mu_1^{\text{meas}}, \ldots, \mu_n^{\text{meas}}]$. For that particular case, the conjugate prior should be chosen to be a Gaussian distribution [14]:

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}) \tag{15}$$

where $\boldsymbol{\Lambda}_0^{-1} = \boldsymbol{\Sigma}_0$. If the mean of the distribution is unknown but the inverse covariance matrices $\boldsymbol{\Lambda}^{\text{state}}$ and $\boldsymbol{\Lambda}^{\text{meas}}$, are unknown, the conjugate prior is Wishart distribution [14]:

$$p(\boldsymbol{\theta}) = W(\boldsymbol{\theta}|\mathbf{W}, \nu) \tag{16}$$

where $\boldsymbol{\theta} = [\lambda_{11}^{\text{state}}, \ldots, \lambda_{nn}^{\text{state}}, \lambda_{11}^{\text{meas}}, \ldots, \lambda_{nn}^{\text{meas}}]$ contains coefficients of $\boldsymbol{\Lambda}^{\text{state}}$ and $\boldsymbol{\Lambda}^{\text{meas}}$, respectively. ν is the number degrees of freedom of the distribution and $\mathbf{W}$ is a scale matrix [14]. Finally, if both the mean and the covariance matrices are unknown, $\boldsymbol{\mu}^{\text{state}}, \boldsymbol{\mu}^{\text{meas}}, \boldsymbol{\Lambda}^{\text{state}}, \boldsymbol{\Lambda}^{\text{meas}}$, the conjugate prior is normal-Wishart distribution [14]:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1}) W(\boldsymbol{\Lambda}|\mathbf{W}, \nu) \tag{17}$$

where β is a constant. As expected with iterative algorithms, the chain has a convergence time denoted by M_{init} and this needs to be taken into account when computing the mean $E[\boldsymbol{\theta}]$. Moreover, the MCMC assumes that the drawn samples $\boldsymbol{\theta}_k$ must be uncorrelated and therefore it is important to monitor the autocorrelation function $R[n]$ and skips some samples if necessary. Finally, it is useful to monitor the acceptance rate as it denotes the number of the accepted proposal points $\boldsymbol{\theta}_m$. For a detailed discussion on convergence criteria see [18].

In Figs. 1–3, an example of the joint parameter and state estimation employing MH algorithm and Kalman filter is illustrated for the following SSM:

$$x_k \sim N(ax_{k-1}, b) \tag{18}$$

$$y_k \sim N(cx_k, d) \tag{19}$$

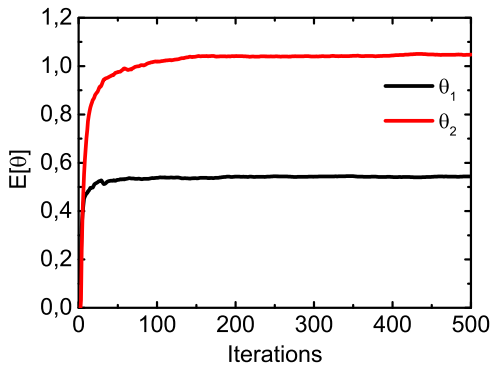

Fig. 3. The estimated mean of the parameters for increasing number of iterations.

where the parameter vector is $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4] \equiv [a, b, c, d]$. The SSM shown in equation (18), (19) can be related to the characterization of an AR process, $\mathbf{x}_k$, (phase noise or nonlinear phase noise) in terms of the coefficient a and variance, b, from the measured data y_k, where d represents measurement noise variance and c is the scaling coefficient. In the following, we would like to estimate the parameters describing the state evolution, i.e., a and b. The task is then to perform estimation of $\boldsymbol{\theta}^{\text{state}} = [a, b]$ by employing the MH algorithm in combination with Kalman filtering as explained in Algorithm 2.

In Figs. 1–3, the output of MCMC algorithm employing MH and Kalman filtering is shown. Figs. 1 and 2 show the histogram of the estimated parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2] \equiv [a, b]$. In Fig. 1, the number of samples is $M = 5000$ and number of lags is $P = 1$ while in Fig. 2, the number of samples is $M = 500$ and number of lags is $P = 10$. Comparing Figs. 1 and 2 illustrates that for the considered case increasing the number of lags and decreasing the number of samples decreases the variance of the posterior distribution of $\boldsymbol{\theta}$. In Fig. 3, the estimated mean of the parameters is plotted as a number of iterations m. Fig. 3 shows that the MH algorithm converges after approximately 100 samples and the estimated mean values are close to the true values of 0.5 and 1.

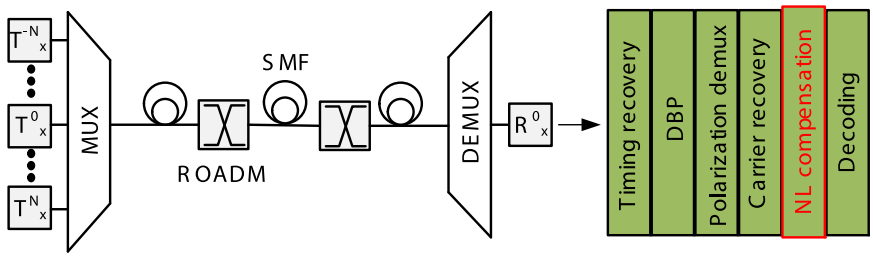

Fig. 4. System set-up for WDM transmission system. MUX: multiplexing, ROADM: reconfigurable add drop multiplexer, DEMUX: demultiplexer, T_x^N : Nth transmitter, SMF: single mode fibre, R_x^0: receiver for the middle channel and NL: nonlinearity.

III. XPM Compensation

In Fig. 4, wavelength division multiplexed system (WDM) is shown together with the digital signal processing (DSP) based demodulation stages. If the single channel digital back propagation (DBP) is used for compensation of intra-channel nonlinearties, it has been shown by Tao *et al.*, [19], that the impact of inter-channel nonlinearities in terms of XPM induced polarization scattering and phase noise, can be approximated by the following closed form channel model:

$$\mathbf{Y}(t) = \mathbf{W}(t)\mathbf{S}(t) + \mathbf{N}(t) \tag{20}$$

where $\mathbf{Y}(t) = [y^x(t), y^y(t)]^T$ and $\mathbf{S}(t) = [s^x(t), s^y(t)]^T$ represent the received and transmitted symbol sequences associated with x and y polarization, and $\mathbf{N}(t) = [n^x(t), n^y(t)]^T$ is the noise. The transfer matrix, $\mathbf{W}(t)$, contains the time-varying XPM effects and is expressed as [19]:

$$\mathbf{W}(t) = \begin{bmatrix} \sqrt{1-|w_{\mathrm{xy}}(t)|^2}e^{j\phi_x(t)} & w_{\mathrm{yx}}(t)e^{j(\phi_x(t)+\phi_y(t))} \\ w_{\mathrm{xy}}e^{j(\phi_x(t)+\phi_y(t))} & \sqrt{1-|w_{\mathrm{yx}}(t)|^2}e^{j\phi_x(t)} \end{bmatrix} \tag{21}$$

where $w_{\mathrm{xy/yx}}(t)$ denotes the effects of XPM induced polarization scattering from x to y polarization and vice versa. The XPM induced phase noise is denoted by $\phi_{x/y}(t)$. Similar model for the nonlinearity transfer function can be deducted from [20]. This can be observed under assumption that if $|w_{\mathrm{xy/yx}}(t)| << 1$ in which case $Y_x(t) = S_x(t)e^{j\phi_x(t)} + w_{\mathrm{yx}}(t)S_y(t)e^{j(\phi_x(t)+\phi_y(t))} = S_x(t)e^{j\phi_x(t)} + \Delta S(t)$, where $\Delta S(t)$ is the perturbation term. The model shown in equation (21) has been a subject for the investigation for different compensation schemes [21], [22]. In most of the cases the full model of $W(t)$ has not been considered. For instance, in [22], the phase terms $\phi_{x/y}(t)$ are set to zero, while in [23], only $\phi_x(t)$ is considered.

It has been shown by simulation and also experimentally that $w_{xy/yx(t)}$, $\phi_x(t)$ and $\phi_y(t)$ can be modeled as a time-varying stochastic processes [19]. In this paper, it is assumed assume that $\mathbf{x}(t) = [w_{\mathrm{xy}}(t), w_{\mathrm{yx}}(t), \phi_x(t), \phi_y(t)]$ can be approximated as an AR process. The AR process is a good approximation for $w_{\mathrm{xy/yx}}(t)$, $\phi_x(t)$ and $\phi_y(t)$ as it can take memory of the system into account. The discrete time AR process, u_k of order M is expressed as:

$$u_k = \sum_{i=1}^{M} a_i u_{k-i} + v[k] \tag{22}$$

where $v_k \sim N(0, \sigma^2_{\mathrm{AR}})$, σ^2_{AR} is the variance of the AR process and $a_1, \ldots, a_M$ are the coefficients stating to what extent the past values influence the time instant k. In order to include the AR process into the state-space framework, equation (22) needs to be expressed as a first order AR process. This can be obtain by introducing the following substitution:

$$\begin{bmatrix} x_k^1 \\ x_k^2 \\ \vdots \\ x_k^M \end{bmatrix} = \begin{bmatrix} u_k \\ u_{k-1} \\ \vdots \\ u_{k-M} \end{bmatrix} \tag{23}$$

and the following is obtained:

$$\overbrace{\begin{bmatrix} x_k^1 \\ x_k^2 \\ \vdots \\ x_k^M \end{bmatrix}}^{\mathbf{x}_k} = \overbrace{\begin{bmatrix} a_1 & \ldots & a_{M-1} & a_M \\ 1 & \ldots & 0 & 0 \\ 0 & \ldots & 1 & 0 \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} x_{k-1}^1 \\ x_{k-1}^2 \\ \vdots \\ x_{k-1}^M \end{bmatrix}}^{\mathbf{x}_{k-1}} + \begin{bmatrix} v_{k-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{24}$$

To be consistent with the probabilistic SSM defined in equation (2), (3), equation (24) is expressed as:

$$\mathbf{x}_k \sim N(\mathbf{x}_k | \mathbf{A}\mathbf{x}_{k-1}, \boldsymbol{\Sigma}) \tag{25}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix except that the coefficient $\boldsymbol{\Sigma}_{11} = \sigma^2_{\mathrm{AR}}$. The order of the AR process will strongly depend on the auto-correlation function describing $\mathbf{x}(t)$ which again will depend on the transmission distance and scenario. To be more specific, the auto-correlation function will depend if the link is dispersion managed or unmanaged, and also what kind of optical amplification is used. Furthermore, in the case of dispersion managed link, the type of dispersion map will play a significant role. Therefore, each situation needs to be treated differently.

In the following, the nonlinear SSM employed for tracking and estimating time-varying parameters describing $\mathbf{W}(t)$ is presented. It is assumed that the effects XPM induced polarization scattering and phase are well approximated by the first order AR model. The state vector, $\mathbf{x}_k$ that is to be estimated is expressed

in discrete time as:

$$\mathbf{x}_k = \begin{bmatrix} x_k^1 \\ x_k^2 \\ x_k^3 \\ x_k^4 \end{bmatrix} = \begin{bmatrix} \Re(w_{xy,k}) \\ \Im(w_{xy,k}) \\ \phi_{x,k} \\ \phi_{y,k} \end{bmatrix} \tag{26}$$

where $\Re$ and $\Im$ denote the real and the imaginary part, respectively. To obtain the values for w_{yx}, the following property is used: $w_{\mathrm{xy}} = -w_{\mathrm{xy}}^*$. The evolution of state vector $\mathbf{x}_k$ is then expressed as:

$$\mathbf{x}_k = \overbrace{\begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{bmatrix}}^{A_w} \mathbf{x}_{k-1} + \begin{bmatrix} v_{k-1}^1 \\ v_{k-1}^2 \\ v_{k-1}^3 \\ v_{k-1}^4 \end{bmatrix} \tag{27}$$

where $v_{k-1}^n, n = 1, .., 4$ describe the process noise and are samples drawn from a Gaussian distribution with a zero mean and variance $\sigma_{w,n}^2$ The observation, measurement, vector needs to be split in real and imaginary part and is expressed as:

$$\mathbf{y}_k = \begin{bmatrix} y_k^1 \\ y_k^2 \\ y_k^3 \\ y_k^4 \end{bmatrix} = \begin{bmatrix} \Re(y_k^x) \\ \Im(y_k^x) \\ \Re(y_k^y) \\ \Re(y_k^y) \end{bmatrix} \tag{28}$$

the transmitted symbols vector and measurement noise are expressed as: $\mathbf{s}_k = [s_k^1, s_k^2, s_k^3, s_k^4]^T = [\Re(s_k^x), \Im(s_k^x), \Re(s_k^y), \Im(s_k^y)]^T$ and $\mathbf{n}_k = [n_k^1, n_k^2, n_k^3, n_k^4]^T = [\Re(n_k^x), \Im(n_k^x), \Re(n_k^y), \Im(n_k^y)]^T$, where T denotes the transpose operation. The measurement equation is then expressed as:

$$\begin{bmatrix} y_k^1 \\ y_k^2 \\ y_k^3 \\ y_k^4 \end{bmatrix} = \begin{bmatrix} h_1(\mathbf{x}_k, \mathbf{s}_k) \\ h_2(\mathbf{x}_k, \mathbf{s}_k) \\ h_3(\mathbf{x}_k, \mathbf{s}_k) \\ h_4(\mathbf{x}_k, \mathbf{s}_k) \end{bmatrix} + \begin{bmatrix} n_k^1 \\ n_k^2 \\ n_k^3 \\ n_k^4 \end{bmatrix} \tag{29}$$

where $\mathbf{H}_k(\mathbf{x}_k, \mathbf{s}_k) = [h_1(\mathbf{x}_k, \mathbf{s}_k), \ldots, h_4(\mathbf{x}_k, \mathbf{s}_k)]^T$ are expressed as:

$$\begin{aligned} h_1(\mathbf{x}_k, \mathbf{s}_k) = & \\ h_2(\mathbf{x}_k, \mathbf{s}_k) = & \\ h_3(\mathbf{x}_k, \mathbf{s}_k) = & \\ h_4(\mathbf{x}_k, \mathbf{s}_k) = & \\ -Bs_k^1(\sin\Delta\phi\cos\phi - \sin\phi\cos\Delta\phi) & \\ +Bs_k^1(\sin\Delta\phi\sin\phi + \cos\Delta\phi\cos\phi) & \\ +Bs_k^3(\sin\Delta\phi\cos\phi - \sin\phi\cos\Delta\phi) & \\ -Bs_k^3(\sin\Delta\phi\sin\phi + \cos\Delta\phi\cos\phi) & \\ -Bs_k^2(\sin\Delta\phi\cos\phi - \sin\phi\cos\Delta\phi) & \\ -Bs_k^2(\sin\Delta\phi\sin\phi + \cos\Delta\phi\cos\phi) & \\ +Bs_k^4(\sin\Delta\phi\cos\phi - \sin\phi\cos\Delta\phi) & \\ +Bs_k^4(\sin\Delta\phi\sin\phi + \cos\Delta\phi\cos\phi) & \\ -s_k^3(w_1\cos\phi - w_2\sin\phi) + s_k^4(w_1\sin\phi - w_2\cos\phi) & \\ -s_k^3(w_1\sin\phi + w_2\cos\phi) - s_k^4(w_1\cos\phi - w_2\sin\phi) & \\ +s_k^1(w_1\cos\phi - w_2\sin\phi) - s_k^2(w_1\sin\phi - w_2\cos\phi) & \\ +s_k^1(w_1\sin\phi + w_2\cos\phi) + s_k^2(w_1\cos\phi - w_2\sin\phi) & \end{aligned} \tag{30}$$

where $B = \sqrt{1 - |x_k^1 + jx_k^2|^2}, \phi = (x_k^3 + x_k^4)/2$ and $\Delta\phi = (x_k^3 - x_k^4)/2$.

In summary, the probabilistic SSM describing the effects of XPM induced polarization scattering and phase noise is expressed as:

$$\mathbf{x}_k \sim N(\mathbf{x}_k | \mathbf{A}_w \mathbf{x}_{k-1}, \boldsymbol{\Sigma}_w) \tag{31}$$

$$\mathbf{y}_k \sim N(\mathbf{y}_k | \mathbf{H}_k(\mathbf{x}_k, \mathbf{s}_k), \boldsymbol{\Sigma}_n) \tag{32}$$

where $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_n$ are process and measurement covariance matrices, respectively. The matrix $\mathbf{H}(\mathbf{x}_k, \mathbf{s}_k)$ is a nonlinear function of the states and iterative estimation of the state vector $\mathbf{x}_k$ is a challenging task as approximative nonlinear filtering methods need to be employed. For estimating the state vector $\mathbf{x}_k$, different types of nonlinear filtering methods need to be applied: extended, unscented and cubature Kalman filter or more general particle filtering. In this paper, extended Kalman and particle filtering are employed. For the implementation of the extended Kalman and particle filter see [16], [15]. At each iteration step, the compensation is performed by applying $\mathbf{W}^{-1}$ to the symbol vector $\mathbf{y}_k$.

In Figs. 5–6, the required optical signal to noise ratio (OSNR) to achieve symbol error rate (SER) of 10^{-3} is computed as a function of increasing variance of XPM induced polarization scattering w_{xy} for dual polarization (DP) QPSK and 16-QAM, respectively. The SER is determined by comparing the transmitted symbol sequence in the two polarizations, $\mathbf{s}_k = [s_k^x, s_k^y]$,

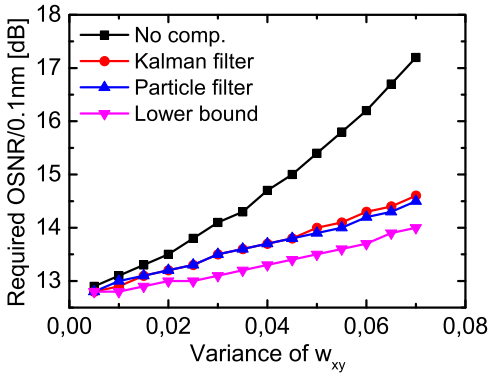

Fig. 5. Required OSNR, to achieve SER of 10^{-3}, as a function of variance of the polarization scattering process, w_{xy}, for 28 Gbaud DP-QPSK system. The XPM induced phase noise is set to zero.

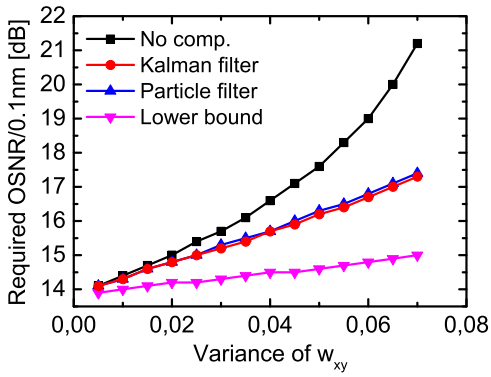

Fig. 6. Required OSNR, to achieve SER of 10^{-3}, as a function of variance of the polarization scattering process, w_{xy}, for 28 Gbaud DP-16QAM system. The XPM induced phase noise is set to zero.

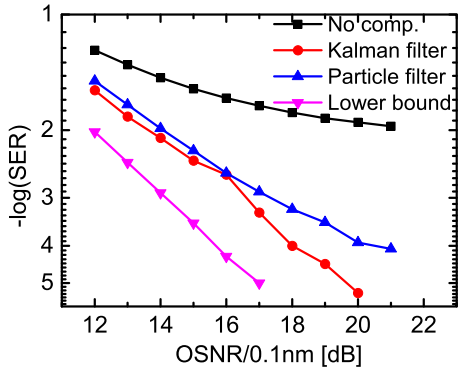

Fig. 7. SER as a function OSNR for 28 Gbaud DP-QPSK in the presence of XPM induced polarization scattering and phase noise with the corresponding variances of 0.03 and 4 deg^2, respectively

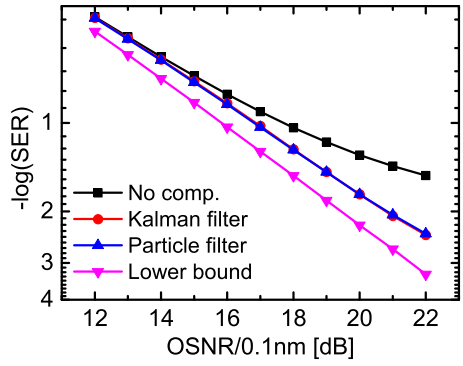

Fig. 8. SER as a function OSNR for 28 Gbaud DP-16QAM in the presence of XPM induced polarization scattering and phase noise with the corresponding variances of 0.01 and 1.71 deg^2, respectively

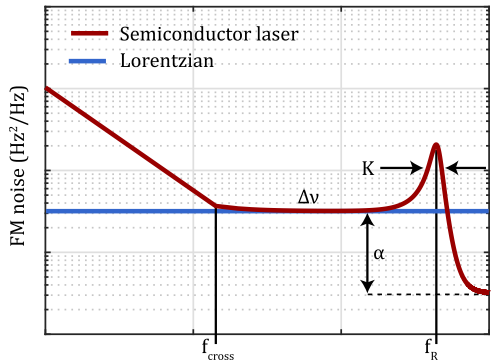

Fig. 9. Power spectral density of FM noise.

to the estimated one obtained thorough the compensation stage $\mathbf{s}_k^{\mathrm{est}} = \mathbf{W}^{-1}\mathbf{y_k}$. The XPM induced phase noise is set to zero. The lower bound is computed by assuming that the process w_{xy} is known at the compensation block and by adding the process noise to it with the variance corresponding to the variance of w_{xy} process. It is observed from Figs. 5–6 that successful compensation of XPM induced polarization scattering is achievable. However, for DP-QPSK we are able to operate closer to the lower bound than for DP-16QAM. The performance of the extended Kalman and particle filter is similar. In Figs. 7–8, the joint impact of XPM induced polarization scattering and phase noise on SER is investigated. The achievable gain in terms of OSNR is approximately 3 dB at SER of $4 \cdot 10^{-2}$ for DP-QPSK and DP-16QAM, respectively.

IV. Carrier Recovery

The majority of carrier recovery algorithms have been designed for sources with Lorentzian lineshape. For more complicated lineshapes, the standard phase recovery algorithm may result in penalties [7]. In Fig. 9, frequency noise (FM) spectra of a laser having Lorentzian and non-Lorentzian lineshape is illustrated. The Lorentzian lineshape is fully characterized by the linewidth $\Delta\nu$ and is expressed by the following stochastic differential equation:

$$\frac{d\phi}{dt} = v(t) \tag{33}$$

where $v(t) \sim N(0, \sigma_\phi^2)$ and σ_ϕ^2 is the variance proportional to the linewidth $\Delta\nu$. The non-Lorentzian lineshape in Fig. 9 is characterized by $1/f$ noise crossing frequency f_{cross}, resonance frequency, f_R, linewidth enhancement factor, α, the width of the resonance peak, K and the linewidth $\Delta\nu$. Rigorously speaking, the stochastic differential equation describing the phase evolution phase of a laser is obtained from the rate equations and is expressed as:

$$\frac{d\phi}{dt} = \alpha G_N N(t) + v(t) \tag{34}$$

where $N(t)$ is the excess carrier concentration and G_N is the differential gain. In general, the laser rate equation are set of continues times stochastic differential equation that can be formulated into discrete state-space framework using Euler-Maruyama discretization scheme [7]:

$$\begin{bmatrix} \Omega_k \\ \phi_k \\ \rho_k \\ N_k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} \Omega_{k-1} \\ \phi_{k-1} \\ \rho_{k-1} \\ N_{k-1} \end{bmatrix} + \begin{bmatrix} v_{k-1}^1 \\ v_{k-1}^2 \\ v_{k-1}^3 \\ v_{k-1}^4 \end{bmatrix}$$

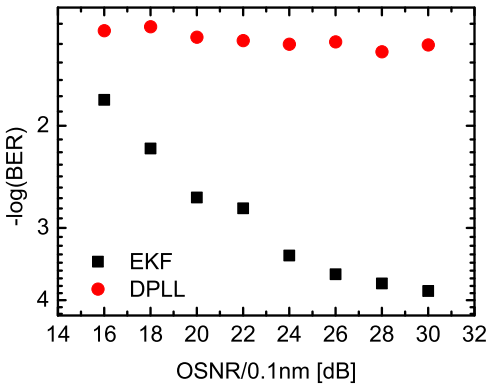

Fig. 10. BER ratio as a function of OSNR for carrier recovery employing digital PLL and rate equations based Kalman filter.

$$\begin{bmatrix} y_k^x \\ y_k^y \end{bmatrix} = \begin{bmatrix} s_k^x e^{i(\Omega_k k T_s + \phi_k)} \\ s_k^y e^{i(\Omega_k k T_s + \phi_k)} \end{bmatrix} + \begin{bmatrix} n_k^1 \\ n_k^2 \end{bmatrix}$$

where Ω_k, ϕ_k, ρ_k, N_k and T_s are frequency offset between transmitter and LO oscillator laser, phase deviation, output intensity perturbation, excess carrier concentration and sampling time, respectively. Coefficients of the matrix $\mathbf{A}$ can be directly related to the rate-equations [7]. We would like to stress that the frequency offset is not an integral part of the rate equations describing the laser. However, due to the coherent intra-dyne detection there will be a non zero frequency offset between the transmitter and the LO laser. If the frequency offset correction is performed before the carrier recovery stage, the frequency offset should not be included in the state-space framework. However, since we are already employing the state-space framework for phase estimation, it is very convenient to include the frequency offset as well. In that way, joint estimation of the frequency and phase is performed. Fig. 10, shows the experiential performance of polarization multiplexed 28 Gbaud 16-QAM, for a semiconductor laser using power spectrum density characterized by a Lorentzian linewidth of 500 kHz, 1 GHz relaxation resonance frequency and 0.1 ns damping factor. The bit error rate (BER) is computed as a function of OSNR for DSP chain employing a standard second-order digital phase locked loop (PLL) and EKF, with SSM described above, for carrier recovery. Employing the PLL data cannot be successfully demodulated as the phase noise model is more complicated than pure Lorentzian. Employing the EKF framework successful data demodulation is achieved.

V. Gaussian Mixture Model

In this section, GMM is briefly reviewed and it is illustrated how GMM can be applied for nonlinear phase noise compensation and to silicon photonics process characterization.

Let us assume that we have acquired a data set $\boldsymbol{\Xi} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ with dimensions $D \times N$, where D is the dimension of a variable $\mathbf{y}$ and N is the number of data points. Superposition of Gaussian distributions can be used to approximate any measured distribution and probability density function of $\boldsymbol{\Xi}$ can thereby be approximated as:

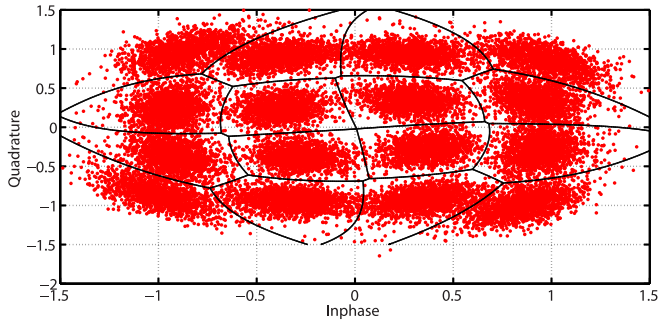

Fig. 11. Experimental results: demodulated signal constellation impaired by nonlinear phase noise for 14 Gbaud DP 16-QAM after 800 km of transmission through a dispersion managed link.

$$p(\boldsymbol{\Xi}) = \sum_{k=1}^{M} \pi_k N(\boldsymbol{\Xi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{35}$$

where M is the number of mixture components, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the mean and the covariance of each component, respectively. Moreover, π_k is a mixing coefficient, $(\sum_k^M \pi_k = 1)$, describing weighting factor of each Gaussian distribution contributing to $p(\boldsymbol{\Xi})$. In order for $p(\boldsymbol{\Xi})$ to properly represent the data set, $\boldsymbol{\Xi}$, the GMM expressed in equation (35) needs to be parameterized in terms of $\pi_k, \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. This is achieved by maximizing the log likelihood function, $\log p(\boldsymbol{\Xi}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the data set $\boldsymbol{\Xi}$. The log likelihood function is expressed as:

$$\ln p(\boldsymbol{\Xi}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{M} \pi_k N(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{36}$$

where N is the number of data points. It is explained in [14] that direct maximization of equation (36) is problematic and can results in significant overfitting. Therefore, the iterative EM framework needs to be used instead to learn the parameters, $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The EM is a two step iterative procedure which is guaranteed to converge to the (local) ML solution. For the implementation of EM algorithm, see [24].

A. Nonlinear Phase Noise Compensation

For the circularly symmetric noise distributions, the noise covariance matrix is diagonal and has equal coefficients. This is typically the case for a linear transmission channel with additive white Gaussian noise. For that particular case, the optimum symbol detection is performed by minimizing the Euclidean distance between a received symbol, $\mathbf{y}_n$ and the corresponding reference constellation. In contrast, for a nonlinear transmission channel dominated by the nonlinear phase noise, Fig. 11, the noise covariance matrix is no longer diagonal with equal coefficients.

The optimal signal detection is then obtained by maximizing *a posteriori* probability of the received symbol $\mathbf{y}_n$ belonging to one of the reference constellation points denoted k, where $k = 1, \ldots, M$ and M is the constellation size:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, p(k|\mathbf{y}_n) \tag{37}$$

or in another words find a constellation point k for which $p(k|\mathbf{x})$ is maximized. The *a posteriori* probability $p(k|\mathbf{y})$ is obtained from Bayes' theorem:

$$p(k|\mathbf{y}_n) = \frac{\pi_k N(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{M} \pi_l N(\mathbf{y}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \tag{38}$$

To compute optimum symbol detection, the posterior probability of a received symbol $\mathbf{y}_n$ belonging to one of the points reference constellation points k needs to be computed:

$$p(k|\mathbf{y}_n) = \arg\max_k \left\{ \frac{\pi_k N(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{M} \pi_l N(\mathbf{y}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \right\} \tag{39}$$

where $n = 1, .., M$ and M is the constellation size, $N(\cdot)$ denotes 2-D Gaussian distribution, $\boldsymbol{\Sigma}_k$ is a 2×2 corresponding covariance matrix, $\boldsymbol{\mu}_k$ are the corresponding means and π_k is a mixing coefficient which is $1/M$ for uniformly distributed constellations. In Eq. (39), it has been assumed that the likelihood function of received symbols has Gaussian distribution. In order to evaluate Eq. (39), parameters of the distribution $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ need to be learned from the received data.

The parameters $\{\Sigma_n, \mu_n\}$ are obtained by solving the optimization problem expressed by equation (39), and is achieved by applying the EM algorithm. In Fig. 11, optimum decision boundaries, in a ML sense, are computed based on equation (39) and displayed for a DP 16-QAM system employing dispersion managed links. It is observed in Fig. 11, that for the considered case, the optimum decision boundaries significantly deviate from the rectangular ones. Nonlinear decision boundaries can also be employed to mitigate nonlinearities and skew associated with in-phase and quadrature modulators.

The EM algorithm needs training data of length N to learn the corresponding means and covariance matrices. During the training period the complexity of the EM algorithm scales as $O(IMN)$, where I is the number of iterations. After the training period (test period), the complexity of the EM algorithm and thereby optimum decision boundaries is reduced to $O(M)$. As the complexity of the EM algorithm scales linearly with the number of training data, N, practical implementation of the algorithm is feasible. The question is what is the sufficient number of the training data points N. This will be dependent on the modulation format as well as the transmission link parameters and is left for the future work.

B. *Learning Integrated Photonics Fabrication Variation*

Nano-scale optical devices such as: directional couplers, integrated microring resonators, nano-cavity based lasers and waveguides are sensitive to the underlying structure variations. Variations in the structure geometry may lead to significant performance degradations and characterizations of such variations is therefore important. Recently, there has been some initial work to predict and quantify the impact of process variations [25]. In many cases, it may prove useful to quantify the underlying distribution of the process variations. Once the distribution has been determined samples can be generated which can potentially be used to determine the impact on the overall system performance.

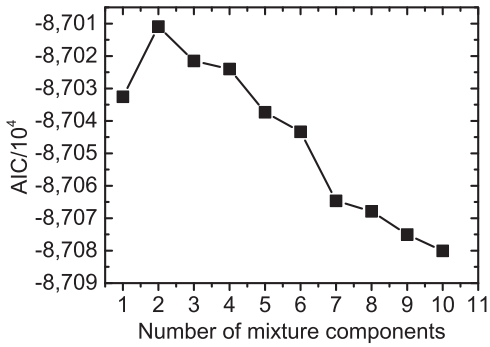

Fig. 12. AIC as a function of number components of GMM.

We follow the example of [25] of SOI based directional coupler were there are variations in the coupler's inner, w_i, and outer, w_e, sidewalls width, i.e., $\mathbf{w} = [w_i, w_o]^T$. It is expected that w_i and w_o are correlated as both are related to the lithography conditions. Next, we form $2 \times N$ data set, $\boldsymbol{\Xi} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$. This example is also closely related to the variations of holes geometry in photonic crystal cavity structures. We would like to determine the underlying distribution of $\boldsymbol{\Xi}$ and therefore approximate $p(\boldsymbol{\Xi})$ with a mixture of Gaussian distributions, equation (35). For the nonlinear phase noise compensation case, the number of mixture components is known, however, for the particular case of SOI coupler, the number of mixture components is unknown. The task is then to estimate the number of mixture components, the subsequent mean vectors and covariance matrices in a ML sense. In order, to find the number of mixture components, maximization of Akaike Information Criterion (AIC) can be used [14]:

$$\text{AIC} = \underset{M}{\text{argmax}} \left\{ \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{M} \pi_k N(\mathbf{w}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} - M \right\}. \tag{40}$$

In short, different models, specified by M, are tried and the one that maximizes AIC is chosen. This is now demonstrated on the example of the SOI coupler. It is assumed that the distribution of $\boldsymbol{\Xi}$ is expressed as in equation (41):

$$p(\boldsymbol{\Xi}) = \sum_{k=1}^{2} \pi_k N(\boldsymbol{\Xi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{41}$$

where $[\pi_1, \pi_2] = [0.6, 0.4]$, $\boldsymbol{\mu}_1 = [9, 6]^T$ nm, $\boldsymbol{\mu}_2 = [8, 7]^T$ nm, $\boldsymbol{\Sigma}_1 = [6, 0; 0, 3]$ nm^2 and $\boldsymbol{\Sigma}_1 = [5, 1; 1, 4]$ nm^2. The task is to approximate the distribution in equation (41) in a best possible way. This can be achieved by using equation (35) and the EM algorithm in combination with AIK criterion. In Fig. 12, AIC is plotted as a function of number of mixing components. It is observed that AIC is maximized when the number of components is two, which is in accordance withe the specified distribution in equation (41). In Figs. 13 and 14, contour plots of the "true" distribution specified by equation (41) and the estimated one by employing EM algorithm for mean and covariance estimation are shown, respectively. It is observed that there is a qualitatively good agreement between Figs. 13 and 14.

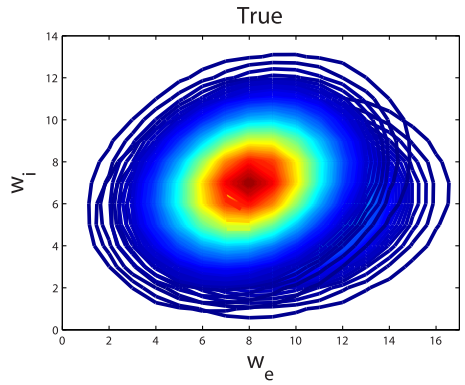

Fig. 13. Contour plot of the 2-D distribution specified by equation (41).

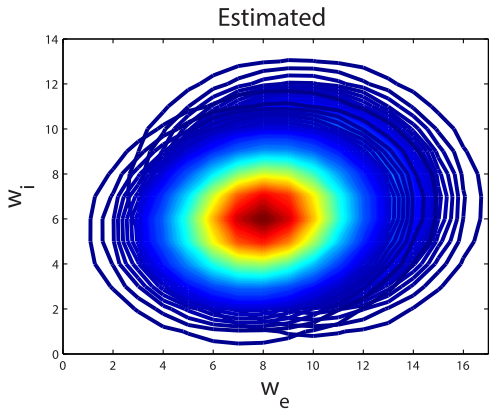

Fig. 14. Contour plot of the estimated 2-D distribution approximating equation (41).

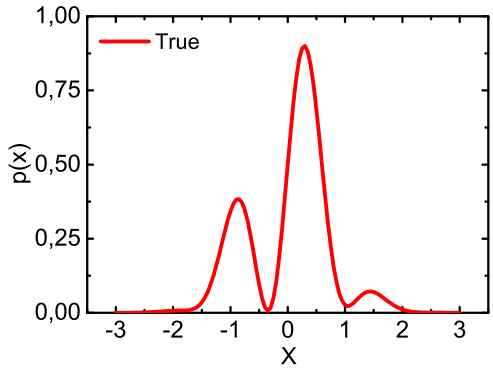

Fig. 15. Probability density function, $p(X)$ of a random variable X.

For evaluating the system performance, one would like to investigate the impact of certain device parameter variations. In this way, it is possible to quantify to which device parameter variations, the system is most sensitive to. For that purpose, we would like to generate samples that have distributions following the device variations. In general, there are many ways of sampling from the certain distribution, see reference [14]. For instance, sampling from the Gaussian distribution can be achieved using the Box-Muller method. However, in some cases if the specified distribution is not Gaussian like the one shown in Fig. 15, it becomes more challenging to perform sampling. For more complex distributions, the MCMC algorithm described in Section II-A can be used to perform sampling. In Fig. 16, histogram of the generated samples using the MH sampling algorithm is shown. It is observed that MH MCMC sampling algorithm is able to generate samples that approximate the distribution specified in Fig. 15.

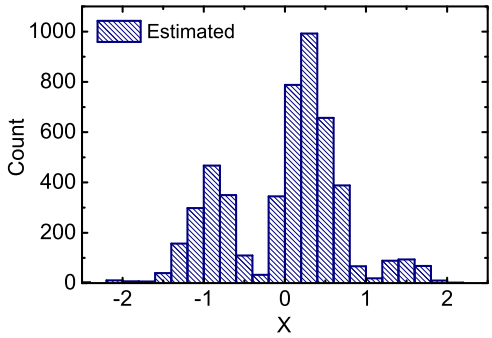

Fig. 16. Histograms of the samples, generated by the MH algorithm, that approximate $p(X)$.

VI. Conclusion

It has been demonstrated that machine learning techniques employing nonlinear state-space framework in combination with MHs sampling algorithm and Bayesian filtering can be employed for a variety of tasks relevant to optical communication such as: system identification, XPM induced polarization scattering and phase noise mitigation and laser rate equation based carrier recovery. The presented framework is a powerful tool as it can be adapted to take into account the underlying physics of the channel and optical components resulting in the overall system improvement. For the considered case of XPM compensation, achievable gains of more than 2 dB have been shown. For the carrier recovery, it has been demonstrated that by employing the laser rate equation based carrier recovery signals with non-Lorentzian lineshape could be demodulated, while employing the second order PLL, (proportional integrator loop filter), signals could not be demodulated. To take the full benefit of the framework, proper characterization and inference of the SSM process equation, describing the state evolution, is crucial. This is especially challenging if the state model is not linear and has high dimensionality, due to induced memory, which may be the case for multi-channel optical transmission systems operating in the nonlinear regime and for the optical sources operating under large signal conditions.

Moreover, it has been demonstrated that the GMM in combination with EM can be employed for computation of nonlinear decision boundaries and characterization of distributions of integrated silicon photonics devices.

Acknowledgment

Research leading to these results has received funding from the Villum Foundation Young Investigator program.

References

[1] J. C. Geyer, C. Doerr, M. Aydinlik, N. Nadarajah, A. Caballero, C. Rasmussen, and B. Mikkelsen, "Practical implementation of higher

order modulation beyond 16-QAM," presented at the European Conf. Exhibition Optical Communication, Los Angeles, CA, USA, 2015, Paper Th1B.1.
[2] T. Richter, S. Member, M. Nölle, F. Frey, and C. Schubert, "Generation and coherent reception of 107-GBd optical Nyquist BPSK, QPSK, and 16QAM," *IEEE Photon. Technol. Lett.*, vol. 26, no. 9, pp. 877–880, May 2014.
[3] G. Raybon, S. Randel, A. Adamiecki, P. J. Winzer, B. Labs, and H.-K. Road, "High symbol rate transmission systems for data rates above 400 Gb/s using ETDM transmitters and receivers," presented at the European Conf. Exhibition Optical Communication, Cannes, Fraance, 2014, Paper Tu.3.3.5.
[4] J. Zhang, J. Yu, Y. Fang, and N. Chi, "High speed all optical Nyquist signal generation and full-band coherent detection." *Sci. Rep.*, vol. 4, pp. 6156-1–6156-8, Jan. 2014.
[5] R. Rios-Müller, J. Renaudier, P. Brindel, H. Mardoyan, P. Jennevé, L. Schmalen, and G. Charlet, "1-Terabit/s net data rate transceiver based on single carrier Nyquist shaped 124 GBaud PDM 32QAM," presented at the Optical Fiber Communication Conf. Exhibition, Los Angeles, CA, USA, 2015, Paper PDP Th5B.1.
[6] C. Laperle and M. O'Sullivan, " Advances in high speed DACs ADCs and DSP for optical coherent transceivers," *J. Lightw. Technol.*, vol. 32, no. 4, pp. 629–643, Feb. 15, 2014.
[7] M. Piels, M. I. Olmedo, S. Member, W. Xue, X. Pang, C. Sch, R. Schatz, G. Jacobsen, I. T. Monroy, S. Member, J. Mork, S. Popov, and D. Zibar, "Laser rate equation-based filtering for carrier recovery in characterization and communication," *J. Lightw. Technol.*, vol. 33, no. 15, pp. 3271–3279, Aug. 1, 2015.
[8] E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *J. Lightw. Technol.*, vol. 32, no. 16, pp. 2862–2876, Aug. 2014.
[9] N. V. Irukulapati, H. Wymeersch, P. Johannisson, and E. Agrell, "Stochastic digital backpropagation," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3956–3967, Nov. 2014.
[10] D. Zibar, M. Piels, R. Jones, and C. G. Schaeffer, "Machine learning techniques in optical communication," presented at the European Conf. Exhibition Optical Communication, Valencia, Spain, 2015, Paper Th.2.6.1.
[11] M. I. Yousefi and F. R. Kschischang, "Information transmission using the nonlinear Fourier transform, Part III: Spectrum modulation," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4346–4369, Jul. 2014.
[12] H. Buelow, "Experimental demonstration of optical signal detection using nonlinear Fourier transform," *J. Lightw. Technol.*, vol. 33, no. 7, pp. 1433–1439, Apr. 1, 2015.
[13] J. E. Prilepsky, S. A. Derevyanko, K. J. Blow, I. Gabitov, and S. K. Turitsyn, "Nonlinear inverse synthesis and eigenvalue division multiplexing in optical fiber channels," *Phys. Rev. Lett.*, vol. 113, no. 1, pp. 13901-1–13901-5, 2014.
[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[15] D. Zibar, L. Henrique, H. D. Carvalho, M. Piels, A. Doberstein, J. Diniz, B. Nebendahl, C. Franciscangelis, J. Estaran, H. Haisch, N. G. Gonzalez, J. C. R. F. D. Oliveira, and I. T. Monroy, "Application of machine learning techniques for amplitude and phase noise characterization," *J. Lightw. Technol.*, vol. 33, no. 7, pp. 1333–1343, Apr. 1, 2015.
[16] S. Sarkka, *Bayesian Filtering and Smoothing*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[17] L. Ljung, "Some classical and some new ideas for identification of linear systems," *J. Control, Autom. Elect. Syst.*, vol. 24, nos. 1/2, pp. 3–10, 2013.
[18] J. C. Spall, "Estimation via Markov chain Monte Carlo," *IEEE Control Syst. Mag.*, vol. 23, no. 2, pp. 34–45, Apr. 2003.
[19] Z. Tao, W. Yan, L. Liu, L. Li, S. Oda, T. Hoshida, and J. C. Rasmussen, "Simple fiber model for determination of XPM effects," *J. Lightw. Technol.*, vol. 29, no. 7, pp. 974–986, Apr. 2011.
[20] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Inter-Channel nonlinear interference noise in WDM Systems: Modeling and mitigation," *J. Lightw. Technol.*, vol. 33, no. 5, pp. 1044–1053, Mar. 2015.
[21] L. Li, Z. Tao, L. Liu, W. Yan, S. Oda, T. Hoshida, and J. Rasmussen, "Nonlinear polarization crosstalk canceller for dual-polarization digital coherent receivers," presented at the Optical Fiber Communication Conf. Exhibition, Los Angeles, CA, USA, 2010, Paper OWE3.
[22] P. Layec, A. Ghazisaeidi, G. Charlet, J.-C. Antona, and S. Bigo, "Generalized maximum likelihood for cross-polarization modulation effects compensation," *J. Lightw. Technol.*, vol. 33, no. 7, pp. 1300–1307, Apr. 2015.
[23] T. Fehenberger, P. M. Yankov, L. Barletta, and N. Hanik, "Compensation of XPM interference by blind tracking of the nonlinear phase in WDM systems with QAM input," presented at the European Conf. Exhibition Optical Communication, Valencia, Spain, 2015, Paper P.5.8.
[24] D. Zibar, O. Winther, and N. Franceschi, "Nonlinear impairment compensation using expectation maximization for dispersion managed and unmanaged PDM 16-QAM transmission," *Opt. Exp.*, vol. 20, no. 26, pp. 181–196, 2012.
[25] T.-W. Weng, Z. Zhang, Z. Su, Y. Marzouk, A. Melloni, and L. Daniel, "Uncertainty quantification of silicon photonic devices with correlated and non-gaussian random parameters," *Opt. Exp.*, vol. 23, no. 4, pp. 4242–4254, Feb. 2015.

Darko Zibar received the M.Sc. degree in telecommunication and the Ph.D. degree in optical communications from the Technical University of Denmark, Lyngby, Denmark, in 2004 and 2007, respectively. He was a Visiting Researcher with the Optoelectronic Research Group, University of California, Santa Barbara, CA, USA, in 2006 and 2008, where he worked on coherent receivers for analog optical links. From February 2009 to July 2009, he was a Visiting Researcher with Nokia-Siemens Networks, where he worked on clock recovery techniques for polarization multiplexed systems. He is currently an Associate Professor at DTU Fotonik, Technical University of Denmark. His research interests include application of machine learning methods to optical communication systems.

Molly Piels received the B.S. degree in engineering, the B.A. degree in history from Swarthmore College in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, Santa Barbara, CA, USA, in 2009 and 2013, respectively. She is currently a Researcher in the High-Speed Optical Communications Group, DTU Fotonik, Lyngby, Denmark. Her research interests include coherent communication and space division multiplexing

Rasmus Jones received the M.Sc. degree in telecommunications engineering from the Technical University of Denmark, Lyngby, Denmark, in 2015. He is currently working toward the Ph.D. degree in fiber-optic communication systems at the Department of Photonics Engineering, Technical University of Denmark. His research interests include digital transmissions and nonlinear signal processing with emphasis on machine learning techniques.

Christian G. Schäffer (M'88) received the Dipl.-Ing. and Dr.-Ing. degrees from the Technical University Berlin, Berlin, Germany, in 1984 and 1989, respectively. From 1988 to 1992, he was with the R&D Center within the Telecommunication Department of DASA, where he was engaged in research on coherent optical generation of microwave signals and optical interconnects for phased array antennas. In 1992, he was with the Fachhochschule Lübeck, Germany, as a Professor with the Department of Electrical Engineering. Between 1999–2009, he was a Full Professor for RF and Photonics, at the Communications Laboratory, Dresden University of Technology, Germany. He is currently with the Helmut Schmidt University, Hamburg, Germany. His research interests include microwave photonics, optical frequency synthesis, Fiber-Bragg gratings, dispersion compensation, Silicon photonics and coherent quantum communication systems.