A comparison of sparse Bayesian regularization methods on computed tomography reconstruction

J Ševčík and V Šmidl
Regional Innovation Center for Electrical Engineering, University of West Bohemia, Pilsen, Czech republic
E-mail: jsevcik@rice.zcu.cz, vsmidl@rice.zcu.cz

Abstract. Design of regularization term is an important part of solution of an ill-posed linear inverse problem. Another important issue is selection of tuning parameters of the regularization term. We address this problem using Bayesian approach which treats tuning parameters as unknowns and estimates them from the data. Specifically, we study a regularization model known as Automatic Relevance Determination (ARD) and several methods of its solution. The first approach is the conventional Variational Bayes method using the symmetrical factorization of the posterior of the vector of unknowns and the vector of tuning parameters. The second approach is based on the idea of marginalization over the vector of unknowns or the vector of tuning parameters, while the complementary vector is estimated using maximum likelihood. The resulting algorithm is thus an optimization task with non-convex objective function, which is solved using standard gradient methods.

The proposed algorithms are tested on real tomographic X-ray data and the comparison with conventional regularization techniques (Tikhonov and Lasso) is performed. The algorithm using marginalization over the tuning parameter is found to be closest to the ground truth with acceptable computational cost.

MATLAB® implementation of the reconstruction algorithms is freely available for download.

1. Introduction
Reconstruction of a tomographic image can be posed as a linear inverse problem. The choice of regularization term for such problem is essential and many possibilities are available, ranging from the basic Tikhonov regularization to total-variation [1]. Motivated by artifacts that are present at reconstructions with low number of projections, see Figure 1, we study performance of regularizations promoting sparsity. The idea is that the white pixels appearing incorrectly in the black area have no support in the data and sparse regularization should suppress them. While Lasso [2] is a prominent example of sparsity regularization, we focus our attention to Bayesian methods based on automatic relevance determination (ARD), which is the main regularization approach used in relevance vector machines [3] or sparse Bayesian learning [4].

Although the ARD principle can be applied to various transformations of the unknown vector, such as the wavelet transformation or first-order difference matrix in image processing [5, 6], we do not impose any transformation and study the effect of the sparsity assumption directly on the vector of unknowns. This is motivated by our aim to suppress artifacts in Figure 1. The probabilistic model of the ARD is based on the assumption of uncorrelated entries of the unknown vector with zero mean and unknown precision. Full Bayesian inference of such model
Ground Truth 20 projections 60 projections 120 projections

Figure 1. Illustration of the artifacts in the images reconstructed from low number of observations obtained using Tikhonov regularization (2) with $\lambda = 1$.

is analytically intractable and methods of approximate inference have to be used. We test three approximations in detail. The first approximation is the common Variational Bayes using factorization of the posterior into the product form (e.g. [7]). The second is the empirical Bayesian approach, which uses analytical marginal over the variable of interest to find best value of the tuning parameters. The third is marginalization over the tuning parameter and maximizing the marginal likelihood of the parameter of interest, which is sometimes referred to as Type II maximum likelihood [8]. In spite of many theoretical results [9], performance on a particular problem is hardly predictable. Therefore, we have implemented three methods of the model and applied them to the problem of tomographic reconstruction of a walnut [10]. Moreover, due to the need to run the methods for high dimensional problem, some standard operations such as inverse of large covariance matrices are prohibitive and are replaced by some common approximations. Therefore, good performance of the tested algorithms that is known from low dimensional problems may be violated.

As an alternative, we also test the hypothesis that the artifacts are caused by different precision of the projections at individual pixels. Relaxing the observation precision yields actually a very similar model to the relaxed penalization of the prior via the ARD model.

2. Problem formulation

In this paper, we will be concerned with a high-dimensional linear inverse problem

$$\mathbf{m} = \mathbf{A}\mathbf{x} + \mathbf{\varepsilon}, \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a known matrix, $\mathbf{x} \in \mathbb{R}^N$ is a vector of unknowns, $\mathbf{m} \in \mathbb{R}^M$ is an observation vector and $\mathbf{\varepsilon} \in \mathbb{R}^M$ is an unknown realization of the noise or an error term.

There is a wide variety of applications that can be described by the mathematical model (1). Although the frameworks introduced in this contribution are fully general, our approaches will be illustrated on the problem of tomographic image reconstruction. From this point of view, vector $\mathbf{m}$ in (1) is composed of pixels of $k$ measured projections (so called sinogram), matrix $\mathbf{A}$ stands for a measurement matrix which describes geometry of projections [11] and vector $\mathbf{x}$ is the reconstructed image in vector form.

The noise term in the computed tomography (CT) data, $\mathbf{\varepsilon}$ in (1), comprises two main sources – an additive (Gaussian) electrical noise and a quantum noise from the photon counts that has a Poisson statistic [12, 13]. However, we will use commonly used Gaussian approximation even for the term $\mathbf{\varepsilon}$ [7, 14, 15]. Since we do not model the noise correlation, we assume that the noise is not correlated. If the noise covariance matrix is known, the problem can be transformed into our formulation (1) by the change of coordinates, $\tilde{\mathbf{m}} = L^{-1}\mathbf{m}$ and $\tilde{\mathbf{A}} = L^{-1}\mathbf{A}$, where $L^{-1}$ is the inverse of the Choleski decomposition of the noise covariance matrix.
Since the problem of tomography reconstruction is typically ill-posed, regularization plays a crucial role. Especially when the number of projections is low, the reconstructed image contains many artifacts, as illustrated in Figure 1. We investigate the use of various regularization terms to suppress the artifacts. The starting point of our research is the most basic Tikhonov regularization

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^N} \{||Ax - m||_2^2 + \lambda||x||_2^2\} ,$$

(2)

where $\lambda$ is a chosen constant. The method is known to provide good results for suitable choice of $\lambda$. While there is a number of recommendations and heuristics for the choice of $\lambda$, systematic estimation of this parameter can be provided by the Bayesian approach. Since the regularization methods based on Bayesian formulation are successfully used in related fields of image processing, we investigate their use in this context. The Bayesian approach treats tuning parameters as unknowns and estimates them from the data. This advantage is obtained at the cost of increased computational requirements of the resulting methods. Since exact solution of the Bayesian inversion problem including tuning parameters estimation is intractable, approximate solutions need to be used.

In this contribution, we compare several possible approximations with respect to accuracy and computational cost of the solution.

3. Bayesian approach

From the Bayesian perspective, the optimization problem (2) can be interpreted as a maximum a posteriori estimate of the following probabilistic model

$$p(m|x, \beta) = \mathcal{N}(m|Ax; \beta^{-1}I_M) ,$$

$$p(x|\alpha) = \mathcal{N}(x|0; \alpha^{-1}I_N) ,$$

(3)

(4)

where $\mathcal{N}$ denotes multivariate Gaussian distribution (defined by (A.1) in the appendix) and $I_M$ denotes identity matrix of size $M \times M$. Parameters $\alpha, \beta$ are chosen constants, representing precision (inverse variance) of the elements of vectors $m$ and $x$, respectively. Logarithm of the joint distribution

$$\log p(m,x|\alpha, \beta) = -\frac{1}{2}\beta||Ax - m||_2^2 - \frac{1}{2}\alpha||x||_2^2 + c$$

(5)

is a sum of the same quadratic forms as in (2) and an additive constant $c$ which aggregates all terms independent of $m, x$. Maximizing (5) is equivalent to minimization of (2) which is obvious when the right hand side of (5) is divided by $-\beta/2$ and assigning $\lambda = \frac{\alpha}{\beta}$. Following the Bayesian approach, the precision parameters can be treated as unknowns and thus estimated from the data. However, even the best manually tuned result of the Tikhonov regularization contains the artifacts. We conjecture that this is a consequence of the common precision parameters for all elements of the unknown vector or the residue. Therefore, we test models where precision parameters $\alpha$ or $\beta$ are not scalars, but vectors where each element models covariance of the corresponding element of vector $m$ or $x$.

We investigate two possible hypotheses on the origin of the artifacts. First, pixels in the resulting image that do not have sufficient support in the inversion problem should have higher penalization towards zero value. This hypothesis corresponds to replacing scalar $\alpha$ by a vector. Second, the error $\epsilon$ is not homogeneous and some of its elements have higher variance than others. This hypothesis corresponds to replacing scalar $\beta$ by a vector. These two hypotheses will be now elaborated into full probabilistic models.

The first hypotheses corresponds to replacing (4) by a different prior model

$$p(x|\alpha) = \mathcal{N}(x|0; \text{diag}(\alpha^{-1})) ,$$

(6)
where $\alpha \in \mathbb{R}^N$ is a vector of precisions for each pixel in the image. This model is known as the Automatic Relevance Determination ARD [3]. It can be shown, that this model assigns relatively high prior probability that the elements of the vector are equal to zero. Strong information in the data is then required to increase the posterior probability of the non-zero values. In effect, estimation of the individual precisions $\alpha_i$ promotes sparse solutions [4], therefore this model will be called sparse prior.

The second hypotheses implies replacement of the likelihood (3) by a model with non-homogeneous error

$$p(m|x, \beta) = N(m|Ax; \text{diag}(\beta^{-1})),$$

where $\beta \in \mathbb{R}^M$ is a vector of precisions for each element of the measured vector $m$. This model will be called non-homogeneous noise model.

Since the vectors $\alpha$ and $\beta$ are considered to be unknown, we define their prior distribution. The conjugate distribution for the precision parameters of a Gaussian distribution is the Gamma distribution [16]. Therefore, we choose the gamma distribution for parameter $\beta$ in homogenous noise prior (3) and for all elements of vector $\alpha$ in the sparse-prior case

$$p(\beta) = \mathcal{G}(\beta|\gamma_0\beta; \delta_0\beta),$$

$$p(\alpha_i) = \mathcal{G}(\alpha_i|\gamma_0\alpha; \delta_0\alpha), \quad i = 1, \ldots, N,$$

where $\gamma_\bullet$ and $\delta_\bullet$ are hyper-prior shaping constants. Priors for vector $\beta$ in the non-homogenous noise prior case and scalar $\alpha$ in the regular prior (4) are:

$$p(\alpha) = \mathcal{G}(\alpha|\gamma_0\alpha; \delta_0\alpha),$$

$$p(\beta_j) = \mathcal{G}(\beta_j|\gamma_0\beta_j; \delta_0\beta_j), \quad j = 1, \ldots, M.$$  

We will study two specific combinations of the proposed models, namely the homogenous noise (3) with sparse prior (6)

$$p(m, x, \alpha, \beta) = p(m|x, \beta)p(x|\alpha)p(\beta)\prod_{i=1}^{N} p(\alpha_i)$$

and the non-homogeneous noise (7) with regular prior (4)

$$p(m, x, \alpha, \beta) = p(m|x, \beta)p(x|\alpha)p(\alpha)\prod_{i=1}^{M} p(\beta)\prod_{j=1}^{M} p(\beta_j).$$

The above models can easily be extended to different versions. For example, replacement of norm $||x||$ in (2) by an operator $||Dx||$ where $D$ may be e.g. a differential operator. It is even possible to consider the elements of the operator to be unknown [17], however its extension to high-dimensions is challenging. Also, it is also possible to combine likelihood (7) with prior (6), we note that the resulting model exhibits extremely high flexibility due to growing number of unknown parameters. The posterior is highly multi-modal and reliable estimation is a challenge. Therefore, we focus only on the simpler models introduced above.

Graphical illustration of the proposed models is given in Figure 2 using the notation of graphical models [18].
Figure 2. Directed graphical model representing the linear inverse problem with Tikhonov regularization (left), sparse prior model (middle), and non-homogeneous noise model (right).

4. Approximate Bayesian Solutions

The optimal inference for a probabilistic model is given by the Bayes rule. In our case, we seek posterior distribution of the unknown vector $\mathbf{x}$ given by measurements $\mathbf{m}$:

$$p(\mathbf{x}|\mathbf{m}) = \frac{p(\mathbf{x}, \mathbf{m})}{p(\mathbf{m})} = \frac{\int_\theta p(\mathbf{x}, \mathbf{m}|\theta)p(\theta)d\theta}{p(\mathbf{m})},$$

(14)

where $\theta$ is a symbol aggregating all unknown parameters and hyperparameters. In our case, $\theta$ is either $\theta = [\alpha, \beta]$ or $\theta = [\alpha, \beta]$. The integration of the right hand side of (14) expresses our aim to make the inference independent of the choice of the tuning parameters.

Since the exact evaluation of (14) for our choice of priors is not analytically tractable, various approximations can be used for obtaining a solution. We will briefly review the most popular principles.

**Marginal likelihood** also known as type II likelihood, is a technique based on evaluating only maximum of the posterior distribution (14),

$$\hat{x}_{ml} = \arg \max_x p(\mathbf{x}|\mathbf{m}).$$

(15)

Since only maximum is required, it is not necessary to evaluate the marginalization constant $p(\mathbf{m})$ in (14) which makes the computation much simpler.

**Empirical Bayes** is a technique for choosing a fixed value of the parameters $\theta$ via

$$p(\mathbf{x}|\mathbf{m}) \approx p(\mathbf{x}|\mathbf{m}, \tilde{\theta}), \quad \tilde{\theta} = \arg \max_\theta \int p(\mathbf{x}, \theta|\mathbf{m})d\mathbf{x},$$

(16)

which can be used if the integration over $\mathbf{x}$ is tractable.

**Variational Bayes** is based on approximation of the intractable joint probability (14) by a tractable approximation (denoted by symbol $q$ to distinguish it from the exact solution) from the class of conditionally independent distributions fulfilling

$$p(\mathbf{x}, \theta|\mathbf{m}) \approx q(\mathbf{x}, \theta|\mathbf{m}) = q(\mathbf{x}|\mathbf{m})q(\theta|\mathbf{m}) = q(\mathbf{x}|\mathbf{m}) \prod_{i=1}^r q(\theta_i|\mathbf{m}),$$

(17)

where the multivariate parameter $\theta$ is partitioned into $r$ sub-vectors $[\theta_1, \ldots, \theta_r]$. The resulting $q(\mathbf{x}|\mathbf{m})$ is then used to approximate (14).
The main result of the Variational Bayes theory is that a conditionally independent approximation closest to the true posterior in the sense of the Kullback-Leibler divergence has to fulfil:

\[ q^*(x|m) \propto \exp \left( E_{q^*(\theta|m)} \left[ \ln (p(x, \theta, m)) \right] \right), \tag{18} \]

\[ q^*(\theta_i|m) \propto \exp \left( E_{q^*(x|m)} \left[ \prod_{j=1, j \neq i} \ln(p(x, \theta, m)) \right] \right), \quad i = 1, \ldots, r, \tag{19} \]

where \( E_{q^*(\theta|m)}[\cdot] \) denotes the expected value of its argument with respect to \( q^*(\theta|m) \) distribution. Since the approximations (18) – (19) are not given in closed form (for a computation of (18) we need to know (19), and vice-versa), an iterative algorithm for finding the solution is usually employed in the VB method.

In the following Sections, we will derive solutions of the considered models and compare their results.

5. Variational Bayes Solutions and Algorithms

In this section, we apply the Variational Bayes method [19] to probabilistic prior models discussed in previous sections and we propose two iterative algorithms suitable for high-dimension problem such as the tomography reconstruction.

First, we show a procedure for the model presuming sparse prior and homogenous noise, i.e. model described by (12).

5.1. Sparse prior

As mentioned earlier, estimation of unknown reconstruction and all unknown parameters (including hyperparameters) can be obtained by the Bayes rule which is in this case computationally intractable. Following variational Bayesian framework we seek a posterior distribution satisfying posterior conditional independence

\[ q(x, \alpha, \beta|m) = q(x|m)q(\beta|m) \prod_{i=1}^{N} q(\alpha_i|m). \tag{20} \]

Application of the main VB result (18)–(19) to model (12) with partitioning (20) yields the approximate marginals in the following form:

\[ q^*(x|m) = N(x|\mu; \Sigma), \tag{21} \]

\[ q^*(\alpha_i|m) = \mathcal{G}(\alpha_i|\gamma_{\alpha_i}; \delta_{\alpha_i}), \quad i = 1, \ldots, N, \tag{22} \]

\[ q^*(\beta|m) = \mathcal{G}(\beta|\gamma_{\beta}; \delta_{\beta}) \tag{23} \]

with shaping parameters given by

\[ \Sigma = \left( A^T \overline{\beta} A + \text{diag}(\alpha) \right)^{-1}, \quad \mu = \Sigma A^T \overline{\beta} m, \tag{24} \]

\[ \gamma_{\alpha_i} = \gamma_0 + \frac{1}{2}, \quad \delta_{\alpha_i} = \delta_0 + \frac{1}{2} \overline{x_i^2}, \quad i = 1, \ldots, N, \tag{25} \]

\[ \gamma_{\beta} = \gamma_0 + \frac{M}{2}, \quad \delta_{\beta} = \delta_0 + \frac{1}{2} (m - Ax)^T (m - Ax), \tag{26} \]
where the symbol \( \hat{\cdot} \) denotes expected value of the argument with respect to its approximate distribution \( q(\cdot) \). For completeness and clarity, we present the form of used moments in (24)–(26):

\[
\text{diag}(\alpha) = \text{diag} \left( \frac{\gamma_1}{\delta_{\alpha_1}}, \ldots, \frac{\gamma_N}{\delta_{\alpha_N}} \right) \\
\hat{\beta} = \frac{\gamma_\beta}{\delta_{\beta}} \\
(m - Ax)^T (m - Ax) = (m - A\mu)^T (m - A\mu) + \text{trace}[A^T A\Sigma],
\]

(29)

\[
\hat{x}_i^2 = \mu_i^2 + \Sigma_{i,i}, \quad i = 1, \ldots, N.
\]

(30)

A candidate solution of the set of implicit equations (24)–(30) can be found iteratively starting from an initial guess. Convergence of such algorithm to a local optimal has been proven e.g. in [20].

However, in high dimensional problems, a naive application of the iterative algorithm is prohibitive due to the necessity of covariance matrix calculation (24). The computation of inverse matrix \( N \times N \), where \( N \) is large, implies high memory and time requirements. For that reason, we employ the conjugate gradient method for computation of \( \mu \) in each iteration. In addition to that, we use only diagonal elements of matrix \( \Sigma^{-1} \) to calculate the variances in (30). This is of course an approximation and it may have negative impact on the method performance.

The resulting high-dimensional VB algorithm for sparse prior and homogenous noise (ARD–VB Algorithm) is summarized in Algorithm 1.

**Algorithm 1 ARD–VB**

1) Initialization:

(i) Initialize constants \( \gamma_{0,\alpha}, \delta_{0,\alpha}, \gamma_{0,\beta}, \delta_{0,\beta} \).

(ii) Initialize hyperparameters \( \alpha = (\alpha_1, \ldots, \alpha_N)^T \) and \( \beta \).

2) Iterate equations (24)–(30) modified for high-dimension problem until convergence criterion is met

(i) Solve set of linear equations (e.g. using conjugate gradient method) for \( x \)

\[
(A^T \beta A + \text{diag}(\alpha)) x = A^T \beta m
\]

(31)

(ii) Invert diagonal elements of matrix \( \Sigma^{-1} \), \( \Sigma_d \) is diagonal matrix,

\[
\text{diag}(\Sigma_d) \leftarrow 1/\text{diag} \left( A^T \beta A + \text{diag}(\alpha) \right)
\]

(32)

(iii) Compute hyperparameter estimates

\[
\alpha_i \leftarrow \frac{2\gamma_{0,\alpha} + 1}{2\delta_{0,\alpha} + x_i^2 + \text{diag}(\Sigma_d)_i}, \quad i = 1, \ldots, N
\]

\[
\hat{\beta} \leftarrow \frac{2\gamma_{0,\beta} + M}{2\delta_{0,\beta} + (m - Ax)^T (m - Ax) + \text{tr}[A^T A\Sigma_d]}
\]

(33)

(34)

Note that expression \( \text{tr}[A^T A\Sigma_d] \) in (34) is equivalent to the inner product \( \langle \text{diag}(A^T A), \text{diag}(\Sigma_d) \rangle \).

The choice of initial values of the shape and scale parameters \( \gamma_\bullet \) and \( \delta_\bullet \) can incorporate prior knowledge about the variances of reconstruction \( x \) and noise \( \varepsilon \). In case we don’t know these
variances, it is appropriate to use values describing non-informative distribution of hyperparameters [19]. Non-informative prior is obtained for prior shape and scale parameters of the Gamma distribution equal to zero, however, for numerical reasons we set them to very low values e.g. $10^{-10}$.

### 5.2. Non-homogeneous noise

Similar derivation can be applied to the model of non-homogeneous noise with regular prior, i.e. model described by (13). In this case, the posterior distribution is restricted to the following posterior conditional independence:

$$q(x, \alpha, \beta|m) = q(x|m)q(\alpha|m)\prod_{j=1}^{M} q(\beta_j|m).$$

(35)

The optimal approximative posterior densities found by the VB method for this model are

$$q^*(x|m) = \mathcal{N}(x; \mu; \Sigma),$$

(36)

$$q^*(\alpha|m) = \mathcal{G}(\alpha|\gamma_\alpha; \delta_\alpha),$$

(37)

$$q^*(\beta_j|m) = \mathcal{G}(\beta_j|\gamma_\beta_j; \delta_\beta_j), \quad j = 1, \ldots, M$$

(38)

where the equations for shaping parameters are shown in the Appendix.

In effect, the algorithm is a repetitive evaluation of the weighted least squares with Tikhonov regularization with adaptive modification of the weighting coefficients. The algorithm is summarized in Algorithm 2 and will be denoted as the WLS–VB algorithm in the remainder of this text.

**Algorithm 2 WLS–VB**

1) Initialization:

   (i) Initialize constants $\gamma_{0\alpha}, \delta_{0\alpha}, \gamma_{0\beta}, \delta_{0\beta}$.
   (ii) Initialize hyperparameters $\alpha$ and $\beta = (\beta_1, \ldots, \beta_M)^T$.

2) Iterate equations (B.1)–(B.7) from Appendix modified for high-dimension problem until convergence criterion is met

   (i) Solve set of linear equations (e.g. using conjugate gradient method) for $x$

   $$(A^T \text{diag}(\beta) A + \alpha I_N) x = A^T \text{diag}(\beta) m$$

   (39)

   (ii) Invert diagonal elements of matrix $\Sigma^{-1}$, $\Sigma_d$ is diagonal matrix,

   $$\text{diag}(\Sigma_d) \leftarrow 1/\text{diag}(A^T \text{diag}(\beta) A + \alpha I_N)$$

   (40)

   (iii) Compute hyperparameter estimates

   $$\alpha \leftarrow \frac{2\gamma_{0\alpha} + N}{2\delta_{0\alpha} + x^T x + \text{tr}(\Sigma_d)}$$

   (41)

   $$\beta_j \leftarrow \frac{2\gamma_{0\beta} + 1}{2\delta_{0\beta} + (m - Ax)^2_j + (A\Sigma_d A^T)_{j,j}} \quad j = 1, \ldots, M$$

   (42)

The adjustments for the high dimensional problem in WLS–VB Algorithm are equivalent to those used in the ARD–VB Algorithm. Recommendations for an initialization of WLS–VB
Algorithm and its properties are the same as the recommendations and properties discussed for the ARD–VB Algorithm.

6. Methods based on analytical marginalization
In this Section, we derive algorithms for the homogenous noise models based on the marginal likelihood (15) and empirical Bayes (16). Both of these approaches are based on analytical marginalization over one variable: the marginal likelihood requires marginalization over $\theta$, and empirical Bayes over $x$.

For feasibility and clarity, we state that joint distributions used in this Section will be of the form:

$$p(m, x|\alpha, \beta) = p(m|x, \beta)p(x|\alpha), \quad (43)$$

$$p(m, x, \alpha|\beta) = p(m|x, \beta)p(x|\alpha)\prod_{i=1}^{N}p(\alpha_i), \quad (44)$$

where individual prior distributions on the right side are distributions (3), (6) and (9). However, precision parameter $\beta$ is assumed to be fixed. Its estimation via the Variational Bayes procedure is straightforward.

6.1. Marginalization over vector of reconstruction
The idea and process introduced in this subsection was published by Wipf and Nagarajan in 2008 [4]. Following the idea of empirical Bayes (16), elements of vector $\alpha$ can be estimated from the data using analytical marginalization of (43) over the vector of unknown reconstruction $x$ and subsequent maximization of the resulting marginal likelihood function. It is equivalent to minimizing the cost function

$$\mathcal{L}(\alpha) := -\ln\int p(m|x, \beta)p(x|\alpha)dx = -\ln p(m|\alpha, \beta) \propto \ln |\Sigma| + m^T\Sigma^{-1}m, \quad (45)$$

where $\Sigma := \beta^{-1}I_M + A\text{diag}(\alpha^{-1})A^T$.

Once some $\tilde{\alpha} = \text{arg min}_\alpha \mathcal{L}(\alpha)$ is computed, the estimation of vector of reconstruction $\tilde{x}$ can be obtained as the posteriori mean

$$\tilde{x} = \text{diag}(\tilde{\alpha}^{-1})A^T\Sigma^{-1}m. \quad (46)$$

Several issues related to this model were observed by Wipf and Nagarajan who also proposed a solution [4]. They define a new variable $z \in \mathbb{R}^N$ and an upper bounding auxiliary cost function $\mathcal{L}(\alpha, z)$, for which it holds that $\mathcal{L}(\alpha, z) \geq \mathcal{L}(\alpha)$. Minimizing this cost function over $\alpha$ for a fixed $z$ can be achieved by solving

$$x \leftarrow \text{arg min}_{x \in \mathbb{R}^N} \left\{ ||m - Ax||_2^2 + \frac{2}{\beta} \sum_{i=1}^{N} z_i^{1/2} |x_i| \right\} \quad (47)$$

and then setting $\alpha_i^{-1} \leftarrow z_i^{-1/2}|x_i|$ for $i = 1, \ldots, N$ [4]. The method is summarized in Algorithm 3 and will be denoted as the ARD–margx Algorithm.

For global convergence analysis see [4]. The initial values of all elements of vector $z$ can be set to 1. Note that the optimization problem (47) can be solved by any convex optimization solver, e.g. using non-linear conjugate gradient method as the Fletcher-Reeves method in our case (see [21]), or by a special purpose Lasso solver. Computation of the optimal $z$ in (48) needs to
Algorithm 3 ARD–margx

1) Initialize vector $z = (z_1, \ldots, z_N)^T$ and fixed $\beta$.

2) Iterate until convergence criterion for $\alpha$ is met:
   
   (i) Solve the minimization problem (47).
   
   (ii) Set $\alpha_i^{-1} \leftarrow z_i^{-1/2}|x_i|, \ i = 1, \ldots, N$.
   
   (iii) Compute the optimal $z$ using
   $$ z \leftarrow \text{diag} \left[ A^T \hat{\Sigma}_\alpha^{-1} A \right] \quad (48) $$

3) Compute estimation of $x$ using (46).

compute matrix $\hat{\Sigma}_x^{-1}$. In the same way as in the previous section, we propose a simplification for high-dimensional problems. Specifically, we use only diagonal elements of matrix $\hat{\Sigma}$ to express the inverse matrix approximation (that is diagonal):

$$ \hat{\Sigma}_x^{-1} = 1/\text{diag} \left( \beta^{-1} I_M + \text{Adiag} (\alpha^{-1}) A^T \right). \quad (49) $$

In the third step of the ARD–margx Algorithm, we use the conjugate gradient method to solve the set of linear equations (46). First, we must solve (for unknown $w$)

$$ w = \hat{\Sigma}^{-1} m = (\beta^{-1} I_M + \text{Adiag} (\alpha^{-1}) A^T)^{-1} m \quad (50) $$

and then set $x \leftarrow \text{diag}(\alpha^{-1}) A^T w$.

6.2. Marginalization over vector of hyperparameters of precision

In previous subsection, we discuss a method based on analytical marginalization over vector of unknown reconstruction $x$. In this Section, we start with analytical marginalization of (44) over the vector of hyperparameters $\alpha$ [3], which models precision in prior distribution (6). Thus, we maximize the marginal likelihood function (15), what is in this case equivalent to minimizing the cost function

$$ L(x) := -\ln \int p(m|x, \beta)p(x|\alpha)p(\alpha) d\alpha = -\ln \left( \mathcal{N}(m|Ax, \beta^{-1} I_M) \prod_{i=1}^N \mathcal{T}(x_i|\nu_i; 0; \sigma_i^2) \right) $$

$$ \propto \frac{1}{2} \beta \| m - Ax \|_2^2 + \sum_{i=1}^N \frac{\nu_i + 1}{2} \ln \left( 1 + \frac{x_i^2}{\nu_i \sigma_i^2} \right), \quad (51) $$

where $\mathcal{T}(x_i|\nu_i; 0; \sigma_i^2)$ is non-standardized Student’s t-distribution (A.3) of element $x_i$ of vector $x$ with mean value $\mu_i = 0$ and with parameters $\nu_i := 2\gamma_0$, and $\sigma_i^2 := \frac{\delta_0}{\gamma_0}$, where $\gamma_0$ and $\delta_0$ are parameters of prior distribution (9).

After substituting original parameters for $\nu_i$ a $\sigma_i^2$ in (51) we can express the cost function and its gradient as

$$ L(x) \propto \frac{1}{2} \beta \| m - Ax \|_2^2 + \sum_{i=1}^N \left( \gamma_0 + \frac{1}{2} \right) \ln \left( 1 + \frac{x_i^2}{2\delta_0} \right), $$

$$ (\nabla L(x))_i \propto (\beta A^T (m - Ax))_i + \frac{(2\gamma_0 + 1)x_i}{2\delta_0} + x_i^2, \ i = 1, \ldots, N. \quad (53) $$
This model is often studied for values of $\gamma_0 = 0$ and $\delta_0 = 0$ which are appropriate for Jeffrey's prior. However, this choice implies many local extremes [9], therefore, we will study this method for a range of tuning parameters $\gamma_0$, $\delta_0$.

Then, the unknown reconstruction $\mathbf{x}$ can be found as argument of the minimum of the cost function, $\tilde{\mathbf{x}} = \arg \min_{\mathbf{x}} \mathcal{L}(\mathbf{x})$. We can find this optimum by using e.g. non-linear conjugate gradient method as the Fletcher-Reeves method [21]. We will denote this approach as the ARD–marga Algorithm in this paper. Initial values $\gamma_0$, and $\delta_0$, can be obtained from the implied moments $E_i$ and $\text{Var}_i$ the meaning of which may be easier to comprehend:

$$E_i := \frac{\gamma_i}{\delta_i} \iff \delta_i = \frac{E_i}{\text{Var}_i}$$
$$\text{Var}_i := \frac{\gamma_i}{\delta_i^2} \iff \gamma_i = \frac{E_i^2}{\text{Var}_i}$$

(54)

7. Experimental results
In this Section, the described algorithms are applied to a real CT measurement data. For testing purpose we will use open X-ray tomographic datasets of a walnut [10] available at www.fips.fi/dataset.php. Our algorithms are available for download from http://www.utia.cas.cz/linear_inversion_methods

We compare the resulting reconstructions of all four proposed approaches and those obtained by the Tikhonov ($\ell_2$) regularized reconstruction (2) and with the Lasso ($\ell_1$) regularized reconstruction, $\tilde{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \{||A\mathbf{x} - \mathbf{m}||_2^2 + \lambda ||\mathbf{x}||_1\}$, where $\lambda > 0$ is a fixed regularized parameter. Note, that the Tikhonov reconstruction can be computed simply by using conjugate gradient method to solve the set of equations, $\tilde{\mathbf{x}} = (A^T A + \lambda I_N)^{-1} A^T \mathbf{m}$. We use the optimized pcg function implemented in MATLAB® for solving the conjugate gradient method. The Lasso regularized reconstruction is computed by "11_1s: A Matlab Solver for Large-Scale $\ell_1$-Regularized Least Squares Problems" implemented by K. Koh, S.-J. Kim and S. Boyd [22].

The mentioned approaches and reconstructions will be discussed especially with respect to their ability to suppress artifacts and with respect to execution time requirements. We also quantify the error of reconstruction via the mean squared error, $\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - x_{GT,i})^2$, and the mean absolute error, $\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |x_i - x_{GT,i}|$. The error is defined as a difference between the estimated image and the ground truth $\hat{x}_{GT}$. The ground truth image is a part of the data set and was obtained by filtered back projection computed from high-number (1200) of projections (available in the data set as well).

Exact quantification of the results is problematic especially with respect to different scales of the result provided by different method. For example, the Lasso method explains the projections using as few non-zero elements of $\mathbf{x}$ as possible. Thus, the estimated pixels have much higher value that those from other methods where contributions from many small pixels are summed up. To visualize this effect in the grayscale image (i.e. with pixels between 0..256), we present the results in two forms. First, the resulting reconstructions are shown in the obtained scale. The scale is set from the ground truth image which minimum is set to zero and maximum to 220. Methods providing estimates of $\mathbf{x}$ with higher values are truncated at the maximum allowed value of 256. The overestimated pixels should thus appear brighter than those in the ground truth image. The results are displayed in Figure 3. Second, the contrast and the brightness of the calculated images are adjusted to better fit the ground truth, i.e. images of the estimated $\mathbf{x}$ are scaled and shifted to minimize the term $||a\mathbf{x} + b - \mathbf{x}_{GT}||_2$, where $a, b \in \mathbb{R}$ are chosen contrast and brightness parameters. The results are displayed in Figure 4.

These results in Figures 3 and 4 visualize the effect of various methods. For example, majority of the pixels in $\mathbf{x}$ from the Lasso method is estimated in the correct scale, Figures 3, however, few pixels are much higher than those in the ground truth. Therefore, the adjustment of the
contrast and the brightness yields almost empty image in Figure 4. The same problem is visible for the results of the ARD-VB method. On the other hand, all pixels in the estimates of the ARD-marg and the WLS-VB methods are closer to zero than those in the ground truth. After adjustment of the contrast and the brightness, the results are closer to the ground truth.

MSE and MAE values of the results of reconstruction of all discussed methods without any scaling or conversion are shown in Table 1 for selected regularization coefficient and in Figure 5 for a range or regularization coefficients. Note that due to the same pixels that cause the problem with visualization scale, the results of the Lasso are deteriorating also in terms of MSE and MAE. The values of MSE and MAE of the results adjusted for contrast and brightness are shown in Figure 6 for a range of regularization coefficients and in Table 2 for selected values.

Table 1. MSE and MAE of the estimated reconstructions from the ground truth depending on the number of projections (size of the known vector $m$), without any scaling or shift of reconstructed images.

| approaches          | 20 projections |                   | 60 projections |                   | 120 projections |                   |
|---------------------|----------------|-------------------|----------------|-------------------|-----------------|-------------------|
|                     | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) |
| Tikhonov reg. ($\lambda = 1$) | 7.025 | 6.182 | 3.416 | 4.124 | 2.741 | 3.706 |
| Lasso reg. ($\lambda = 10^{-2}$) | 175.125 | 13.216 | 62.338 | 10.955 | 36.555 | 10.048 |
| ARD–VB Alg. | 98.411 | 7.769 | 148.595 | 14.373 | 64.241 | 11.960 |
| WLS–VB Alg. | 9.820 | 6.940 | 5.189 | 5.061 | 3.553 | 3.830 |
| ARD–marg | 15.620 | 7.243 | 7.778 | 5.061 | 6.139 | 4.605 |
| ARD–marg $\alpha$ ($E_i = 10^{-2}$) | 6.944 | 6.221 | 3.210 | 4.021 | 2.168 | 3.098 |

Table 2. MSE and MAE of estimated reconstructions from the ground truth after adjustment of the contrast and the brightness depending on the number of projections (size of the known vector $m$).

| approaches          | 20 projections |                   | 60 projections |                   | 120 projections |                   |
|---------------------|----------------|-------------------|----------------|-------------------|-----------------|-------------------|
|                     | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) | MSE ($\times 10^{-5}$) | MAE ($\times 10^{-3}$) |
| Tikhonov reg. ($\lambda = 1$) | 6.955 | 6.206 | 3.395 | 4.130 | 2.739 | 3.710 |
| Lasso reg. ($\lambda = 10^{-2}$) | 18.610 | 10.639 | 15.951 | 9.378 | 13.810 | 8.399 |
| ARD–VB Alg. | 17.968 | 10.384 | 18.122 | 10.396 | 15.920 | 9.327 |
| WLS–VB Alg. | 8.744 | 6.846 | 5.146 | 4.951 | 3.236 | 3.772 |
| ARD–marg $\alpha$ ($E_i = 10^{-2}$) | 10.695 | 6.763 | 5.783 | 4.655 | 4.664 | 4.319 |
| ARD–marg $\alpha$ ($E_i = 10^{-2}$) | 6.938 | 6.206 | 3.205 | 4.011 | 2.166 | 3.104 |
Figure 3. Resulting reconstructions of the walnut truncated to the interval [0; 256] where the ground truth image scale is [0; 220] obtained by: the Tikhonov regularization ($\lambda = 1$), the Lasso regularization ($\lambda = 10^{-2}$), ARD–VB Algorithm, WLS–VB Algorithm, ARD–margx Algorithm and ARD–marga Algorithm ($E_i = 10^{-2}$, $\text{Var}_i = 10^{-4}$) from 20, 60 and 120 projections.
Figure 4. Resulting reconstructions of the walnut with the adjusted contrast and brightness and truncated to the interval $[0; 256]$ where the ground truth image scale is $[0; 220]$ obtained by: the Tikhonov regularization ($\lambda = 1$), the Lasso regularization ($\lambda = 10^{-2}$), ARD–VB Algorithm, WLS–VB Algorithm, ARD–margx Algorithm and ARD–marg$\alpha$ Algorithm ($E_i = 10^{-2}$, $\text{Var}_i = 10^{-4}$) from 20, 60 and 120 projections.
Table 3. Time requirements [s] of the tested algorithms. The algorithms have been implemented in MATLAB® (v. R2015b) and computed on a PC (Windows® 7, Intel® Core™ i5-2410M CPU 2.30 GHz, RAM 4 GB) with the initialization used in the provided scripts.

| approaches                  | 20 projections | 60 projections | 120 projections |
|-----------------------------|----------------|----------------|-----------------|
| Tikhonov reg. (λ = 1)       | 0.43           | 0.98           | 1.71            |
| Lasso reg. (λ = 10⁻²)       | 467.50         | 451.78         | 465.73          |
| ARD–VB Alg.                 | 68.82          | 211.34         | 393.90          |
| WLS–VB Alg.                 | 53.72          | 169.95         | 366.72          |
| ARD–margα (Eᵢ = 10⁻²)       | 184.13         | 445.57         | 879.55          |
| ARD–margα (Eᵢ, Varᵢ = 10⁻²) | 11.26          | 30.75          | 72.97           |

Figure 5. Mean squared error (MSE) and mean absolute errors (MAE) of reconstructions of the slice of the walnut without any scaling or shift obtained from 60 projections by the Tikhonov regularization, the Lasso regularization and ARD–margα Algorithm depending on the regularization parameter λ or prior moments Eᵢ and Varᵢ.

Figure 6. Mean squared error (MSE) and mean absolute errors (MAE) of reconstructions of the slice of the walnut with the adjusted contrast and brightness obtained from 60 projections by the Tikhonov regularization, the Lasso regularization and ARD–margα Algorithm depending on the regularization parameter λ or prior moments Eᵢ and Varᵢ.

The best methods with respect to artifact suppression are the ARD–margα algorithm and the WLS–ARD algorithm. However, these algorithms also seems to suppress the signal since the difference from the ground truth is greater than that of the Tikhonov regularization. The only algorithm achieving better MSE and MAE is the ARD–margα algorithm for the reported tuning parameters.
On the other hand, methods using unscaled \( \ell_1 \) norm such as Lasso do not provide good estimated on this data. Similar results are obtained by the ARD-VB algorithm which seems to suffer from simplifications of the covariance matrix that were introduced to achieve computational feasibility in these dimensions.

The best method in our experiments turns out to be the ARD–marg\( \alpha \) method. It achieves marginally better MSE and MAE than the best tuned Tikhonov method with much lower effort in manual tuning. However, at the cost of increased computational cost (Table 3)). We conjecture that this good performance is due to the fact, that it is the only method from the Bayesian approaches that does not require any additional approximations (such as approximate matrix inversions).

8. Conclusion
The focus of this paper is on comparison of various solutions of the hierarchical Bayesian model known as ARD for a linear inverse problem. The studied solutions are based on different approximations and provide remarkably different results on the testing data. Due to the high dimensionality of the testing problem, additional simplifications had to be made in original algorithms to achieve computational feasibility. The only exception was the algorithm based on analytical marginalization over the vector of prior precisions, which also provides results closest to the ground truth. Moreover, it had also the lowest execution time of all algorithms using ARD. The algorithm using non-homogeneous noise model has some promising features, such as artifact supression, but its error is worse than that of the basic Tikhonov regularization. Perhaps a combination of the non-homogeneous noise with the ARD prior on the reconstructed vector may yield an improved solution. However, a more restricted prior parametrization of such model will be needed to cope with multi-modality of the resulting posterior.

Acknowledgments
This research has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the RICE New Technologies and Concepts for Smart Industrial Systems, project No. LO1607.

Appendix A. Probability density function
Probability density function (PDF) of normal (Gaussian) distribution, denoted as \( \mathcal{N} \), of \( N \)-dimensional random vector \( \mathbf{x} \) is defined as

\[
\mathcal{N}(\mathbf{x}|\mu; \Sigma) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right).
\] (A.1)

PDF of gamma distribution, denoted as \( \mathcal{G} \), of random variable \( x \) is defined as

\[
\mathcal{G}(x|\gamma; \delta) = \frac{\delta^\gamma}{\Gamma(\gamma)} x^{\gamma-1} \exp(-x\delta),
\] (A.2)

where \( \Gamma(\gamma) \) is the gamma function evaluated at \( \gamma \). PDF of (non-standardized) Student’s t-distribution, denoted as \( \mathcal{T} \), of random variable \( x \) is defined as

\[
\mathcal{T}(x|\nu; \mu; \sigma^2) = \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right) \sqrt{\pi \nu \sigma^2}} \left( 1 + \frac{(x - \mu)^2}{\nu \sigma^2} \right)^{-\frac{\nu+1}{2}},
\] (A.3)

where \( \mu \in \mathbb{R} \) is mean value, \( \nu > 0 \) expresses number of degrees of freedom and \( \sigma^2 > 0 \) is a scale parameter.
Appendix B. Shaping parameters of the WLS-VB algorithm

Shaping parameters of posterior distributions (36)–(38) were recognized as

$$\Sigma = \left( A^T \tilde{\text{diag}}(\beta) A + \hat{\alpha} I_N \right)^{-1}, \quad \mu = \Sigma A^T \tilde{\text{diag}}(\beta) \mathbf{m}, \quad (B.1)$$

$$\gamma\alpha = \gamma_0\alpha + \frac{N}{2}, \quad \delta\alpha = \delta_0\alpha + \frac{1}{2} \mathbf{x}^T \mathbf{x}, \quad (B.2)$$

$$\gamma\beta = \gamma_0\beta + \frac{1}{2}, \quad \delta\beta = \delta_0\beta + \frac{1}{2} \left( \mathbf{m} - A\mathbf{x} \right)^2, \quad i = 1, \ldots, M. \quad (B.3)$$

where relevant moments are

$$\tilde{\text{diag}}(\beta) = \text{diag} \left( \begin{bmatrix} \gamma_{\beta_1} & \ldots & \gamma_{\beta_M} \\ \delta_{\beta_1} & \ldots & \delta_{\beta_M} \end{bmatrix} \right), \quad (B.4)$$

$$\hat{\alpha} = \frac{\gamma\alpha}{\delta\alpha}, \quad (B.5)$$

$$\mathbf{x}^T \mathbf{x} = \mu^T \mu + \text{trace}[\Sigma], \quad (B.6)$$

$$\left( \mathbf{m} - A\mathbf{x} \right)_i^2 = \left( \mathbf{m} - A\mu \right)_i^2 + \left( \Lambda \Sigma A^T \right)_{i,i}, \quad i = 1, \ldots, M. \quad (B.7)$$

References

[1] Hämäläinen K, Harhanen L, Hauptmann A, Kallonen A, Niemi E and Siltanen S 2014 International Journal of Tomography & Simulation 25 1–25

[2] Tibshirani R 1996 Journal of the Royal Statistical Society. Series B (Methodological) 267–288

[3] Tipping M E 2001 Journal of Machine Learning Research 1 211–244

[4] Wipf D P and Nagarajan S S 2008 Advances in Neural Information Processing Systems 20 1625–1632

[5] Wang L, Mohammad-Djafari A and Gac N 2017 AIP Conference Proceedings vol 1853 (AIP Publishing) p 060003

[6] Dumitru M, Li W, Gac N and Mohammad-Djafari A 2017 2017 IEEE International Conference on Image Processing

[7] Wang L, Mohammad-Djafari A, Gac N and Dumitru M 2016 Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (IEEE) pp 883–887

[8] Wipf D P, Rao B D and Nagarajan S 2011 IEEE Transactions on Information Theory 57 6236–6259

[9] Buchgraber T 2013 Variational sparse Bayesian learning: Centralized and distributed processing Ph.D. thesis Technical University of Denmark

[10] Hämäläinen K, Harhanen L, Kallonen A, Kujanpää A, Niemi E and Siltanen S 2015 (Preprint arXiv: 1502.04064)

[11] Mueller J L and Siltanen S 2012 Linear and Nonlinear Inverse Problems with Practical Applications (Philadelphia: Society for Industrial & Applied Mathematics)

[12] Schulte R, Heil U, Groß D, Bruehlmann D, Dranischnikow E, Schwanecke U and Schoemer E 2011 Dentomaxillofacial Radiology 40 265–273

[13] Tu S J, Shaw C C and Chen L 2006 Physics in medicine and biology 51 1283

[14] Hamalainen K, Kallonen A, Kolehmainen V, Lassas M, Niinimaki K and Siltanen S 2013 SIAM Journal on Scientific Computing 35 B644–B665

[15] Charbonnier P, Blanc-Féraud L, Aubert G and Barlaud M 1997 IEEE Transactions on Image Processing 6 298–311

[16] Robert C P 2007 The Bayesian choice: from decision-theoretic foundations to computational implementation 2nd ed (New York: Springer)

[17] Tichý O, Šmíd V, Hofman R and Stohl A 2016 Geoscientific Model Development 9 4297

[18] Bishop C 2006 Pattern recognition and machine learning vol 1 (Springer New York)

[19] Šmíd V and Quinn A 2006 The variational Bayes method in signal processing (Berlin: Springer)

[20] Sato M 2001 Neural Computation 13 1649–1681

[21] Nocedal J and Wright S (eds) 2006 Numerical Optimization 2nd ed (New York: Springer)

[22] Kim S J, Koh K, Lustig M, Boyd S and Gorinevsky D 2007 IEEE journal of selected topics in signal processing 1 606–617