Linear inverse problem with range prior on correlations and its Variational Bayes inference

Ondřej Tichý and Václav Šmídl

Institute of Information Theory and Automation, Pod Vodarenskou vezi 4, Prague, Czech republic, {otichy,smidl}@utia.cas.cz

Abstract. The choice of regularization for an ill-conditioned linear inverse problem has significant impact on the resulting estimates. We consider a linear inverse model with on the solution in the form of zero mean Gaussian prior and with covariance matrix represented in modified Cholesky form. Elements of the covariance are considered as hyper-parameters with truncated Gaussian prior. The truncation points are obtained from expert judgment as range on correlations of selected elements of the solution. This model is motivated by estimation of mixture of radionuclides from gamma dose rate measurements under the prior knowledge on range of their ratios. Since we aim at high dimensional problems, we use the Variational Bayes inference procedure to derive approximate inference of the model. The method is illustrated and compared on a simple example and on more realistic 6 hours long release of mixture of 3 radionuclides.

1 Introduction

Linear inverse problems are fundamental in many areas of science, signal processing, or machine learning. The conventional least squares method fails when the problem is ill-conditioned. In these cases, appropriate regularizations are beneficial to obtain desirable solution. Most commonly used regularizations are the Tikhonov [3] and LASSO [12] where different norms of the unknown vector are used, l_2 and l_1 respectively.

Both of these methods have Bayesian interpretation with different prior distribution of the unknown vector. However, parameters of these prior distributions are assumed to be known. More flexible models allow for estimation of the hyper-parameters, e.g. in the form of diagonal elements of the prior covariance matrix, which is known as the automatic relevance determination principle [14] since it favors sparse solutions. Theoretically, full covariance matrix can be also estimated using Wishart distribution [13,6]. However, the problem is then over-parametrized and the influence of additional regularization is significant. In this contribution, we are concerned with models where some elements of the covariance matrix are vaguely known and need to be estimated from the data. We assume the knowledge of ranges of selected elements of the covariance matrix. We follow idea of Daniels and Pourahmadi [2] where modified Cholesky decomposition of the covariance matrix is used for longitudinal data. In our model, we restricted the possible interval for specific elements of the covariance matrix using truncated Gaussian distribution. These intervals are expert information and are considered as input of our algorithm.

The proposed approach is illustrated on simple synthetic example where comparison with Tikhonov and LASSO regularizations will be given. In addition, we apply the resulting algorithm on a problem of determination of the source term of an atmospheric release of radiation where ratios of the released nuclides are vaguely known. This scenario is relevant to the case of the Fukushima Dai-ichi nuclear power plant accident [8]. We aim for estimation of the time profile of the release using gamma dose rate (GDR) measurements, so our measurement vector does not contain nuclide-specific concentration activity measurements but bulk gamma dose rates from a mixture of nuclides. Particularly important are prior assumptions on the nuclide ratios and their treatment. These can be obtained, e.g, from physical analysis of the power plant state (reactor inventory combined with assumptions on the accident type) or from a few available nuclide-specific activity concentration samples downwind the release. In our simulated scenario, 6 hours release of a mixture of 3 nuclides is considered and Austria monitoring network is used together with realistic meteorological data.

2 Mathematical Method

We study the following linear inverse problem

$$\mathbf{y} = M\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbf{R}^{p\times 1}$ is vector of measurements corrupted by error vector $\mathbf{e}$ of the same size, $M \in \mathbf{R}^{p\times n}$ is known matrix, and $\mathbf{x} \in \mathbf{R}^{n\times 1}$ is the unknown vector to be estimated. Solution of the noise-less problem via ordinary least square method is $\mathbf{x} = (M^T M)^{-1} M^T \mathbf{y}$, which is often infeasible due to ill-conditioned matrix M.

The problem is typically recast as an optimization problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ ||\mathbf{y} - M\mathbf{x}||_2^2 + \alpha g(\mathbf{x}) \right\}, \tag{2}$$

where $g(\mathbf{x})$ is a regularization term and α is its weight. Common regularization terms are Tikhonov regularization [3] or LASSO regularization [12]:

$$g_{\text{Tikhonov}}(\mathbf{x}) = ||\mathbf{x}||_2^2, \qquad g_{\text{LASSO}}(\mathbf{x}) = ||\mathbf{x}||_1, \tag{3}$$

however, the parameter α needs to be carefully selected or determined. The optimization approach (2) can be interpreted as a maximum a posteriori estimate of a Bayesian model. Many detailed analysis of Bayesian interpretations and also extensions are available, e.g. [7]. For the purpose of this text, we only note that the Tikhonov regularization is equivalent to MAP estimation of probabilistic model

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ -\log p(\mathbf{y}|M, \mathbf{x}) - \log p(\mathbf{x}|\alpha) \right\}, \tag{4}$$

with

$$p(\mathbf{y}|M,\mathbf{x}) = \mathcal{N}_{\mathbf{y}}\left(M\mathbf{x}, I_p\right), \qquad p(\mathbf{x}|\alpha) = \mathcal{N}_{\mathbf{x}}\left(\mathbf{0}, \alpha^{-1}I_n\right), \tag{5}$$

where $\mathcal{N}$ denotes Gaussian distribution and I_p denotes identity matrix with given size. For given α, the Bayesian model is fully equivalent to the optimization problem (2). However, the unknown parameters, α in this case, can be modeled using hierarchical priors and estimated within the model [1].

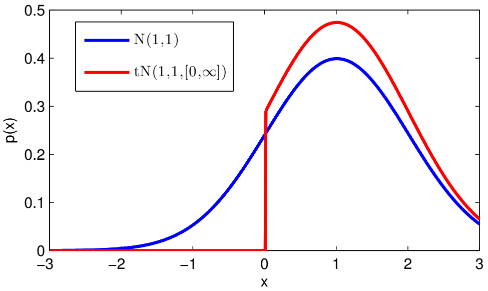

Fig. 1. Example of the Gaussian distribution $\mathcal{N}(1,1)$, blue line, and the truncated Gaussian distribution $t\mathcal{N}(1,1,[0,\infty])$, red line.

For problem specific tasks where assumption on same parameters arise such as non-negativity of $\mathbf{x}$, the optimization approach (2) can be supplemented using "subject to" condition. In Bayesian formulation, this condition can be enforced using truncated Gaussian prior denoted as $t\mathcal{N}$, see one dimensional example in Fig. 1 and Appendix for details.

2.1 Bayesian Hierarchical Model

Consider probabilistic formulation of linear inverse problem (1) with isotropic Gaussian noise

$$p(\mathbf{y}|\mathbf{x},\omega) = \mathcal{N}_{\mathbf{y}}\left(M\mathbf{x}, \omega^{-1}I_p\right), \tag{6}$$

where ω is precision of noise. For unknown ω, we assume prior model in the form of Gamma $\mathcal{G}_\omega(\vartheta_0, \rho_0)$. All prior parameters (subscripted by 0) are set to non-informative values of 10^{-10}. We assume the unknown vector $\mathbf{x}$ to have Gaussian prior; however, with truncated support to positive values,

$$p(\mathbf{x}|\Omega) = t\mathcal{N}_{\mathbf{x}}\left(\mathbf{0}, \Omega^{-1}, [0,+\infty]\right). \tag{7}$$

We aim to model the precision matrix Ω in more detail; hence, we assume Ω in the form of modified Cholesky decomposition as

$$\Omega = L\Upsilon L^T, \tag{8}$$

where $\varUpsilon$ is diagonal matrix with diagonal entries $\boldsymbol{\upsilon} = [\upsilon_1, \ldots, \upsilon_n]$ with prior Gamma model $\mathcal{G}_{\upsilon_j}(\alpha_0, \beta_0)$ for each element and L is lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{2,1} & 1 & 0 & 0 \\ \vdots & \ddots & 1 & 0 \\ l_{n,1} & \ldots & l_{n,n-1} & 1 \end{pmatrix}, \tag{9}$$

with unknown off-diagonal elements forming column vectors $\mathbf{l}_i = [l_{i+1,i}, l_{i+2,i}, \ldots, l_{n,i}]^T \in \mathbf{R}^{(n-i)\times 1}$ for $i = 1, \ldots n-1$. We will introduce prior model for vectors $\mathbf{l}_i$ whose estimates together with estimate of vector $\boldsymbol{\upsilon}$ fully determine the covariance matrix decomposition (8). The prior model for each non-zero element of L, $l_{i,k}$, are chosen as

$$p\left(l_{i,k}|\psi_{i,k}\right) = t\mathcal{N}_{l_{i,k}}\left(0, \psi_{i,k}^{-1}, [a_{i,k}, b_{i,k}]\right), \tag{10}$$

where $\psi_{i,k}$ is unknown precision parameter with prior Gamma model $\mathcal{G}_{\psi_{i,k}}(\zeta_0, \eta_0)$ and with selected interval $[a_{i,k}, b_{i,k}]$ of truncated Gaussian distribution. These intervals allow us to select boundaries for each element of the covariance matrix.

Estimation of the model parameters is analytically intractable; hence, we employ the Variational Bayes method [10] to yield an approximate solution. The Variational Bayes method estimates the posterior solution in the form of conditionally independent distributions that minimize the Kullback-Leibler divergence to the true posterior. This minimization leads to a set of implicit equations which have to be solved iteratively. Here, shaping parameters of recognized posterior distributions

$$\tilde{p}(\mathbf{x}|\mathbf{y}) = t\mathcal{N}_{\mathbf{x}}\left(\mu_{\mathbf{x}}, \varSigma_{\mathbf{x}}, [0, +\infty]\right), \tag{11}$$

$$\tilde{p}(\upsilon_j|\mathbf{y}) = \mathcal{G}_{\upsilon_j}\left(\alpha_j, \beta_j\right), \tag{12}$$

$$\tilde{p}(l_{i,k}|\mathbf{y}) = t\mathcal{N}_{l_{i,k}}\left(\mu_{l_{i,k}}, \varSigma_{l_{i,k}}, [a_{i,k}, b_{i,k}]\right), \tag{13}$$

$$\tilde{p}(\psi_{i,k}|\mathbf{y}) = \mathcal{G}_{\psi_{i,k}}\left(\zeta_{i,k}, \eta_{i,k}\right) \tag{14}$$

$$\tilde{p}(\omega|\mathbf{y}) = \mathcal{G}_{\omega}\left(\vartheta, \rho\right), \tag{15}$$

are iteratively evaluated, see Algorithm 1. The algorithm will be denoted as the least square with the prior adaptive covariance with interval restrictions (LS-APCi) algorithm.

3 Experiments

To test and compare the studied LS-APCi algorithm, we first design a simple synthetic dataset. Second, we perform experiment on realistic gamma dose rate measurement with vaguely known ratios of selected radionuclides.

3.1 Toy Example

We select an ill-conditioned matrix $M \in \mathbf{R}^{6\times 3}$ with elements within 0 and 1 with eigenvalues $[2 \times 10^{-7}, 0.19, 0.23]$. The original vector $\mathbf{x}$ is selected as

Algorithm 1 The least square with the prior adaptive covariance with interval restrictions (LS-APCi) algorithm.

```
1. Initialization
   (a) Set all prior parameters (subscripted by 0) to $10^{-10}$.
   (b) Set initial values: $\langle L\rangle = \langle \Upsilon\rangle = I_n$ and $\langle\omega\rangle = \frac{1}{\max(M^T M)}$.
2. Iterate until convergence or maximum number of iteration is reached:
   (a) Compute moments of $\langle \mathbf{x}\rangle$ using Appendix and shaping parameters of (11):
         $\Sigma_{\mathbf{x}} = \left(\langle\omega\rangle M^T M + \left\langle L\Upsilon L^T\right\rangle\right)^{-1},$   (16)
         $\mu_{\mathbf{x}} = \Sigma_{\mathbf{x}}\left(\langle\omega\rangle M^T \mathbf{y}\right),$   (17)
   (b) Compute moment $\langle\Upsilon\rangle$ using shaping parameters of (12):
         $\boldsymbol{\alpha} = \alpha_0 + \frac{1}{2}\mathbf{1}_{n,1}, \quad \boldsymbol{\beta} = \beta_0 + \frac{1}{2}\mathrm{diag}\left(\left\langle L^T \mathbf{x}\mathbf{x}^T L\right\rangle\right),$   (18)
   (c) Compute moments of $\langle L\rangle$ with restricted ranges using Appendix and shaping
       parameters of (13):
         $\Sigma_{l_{i,k}} = \left(\langle\upsilon_i\rangle\left\langle x_{(i+1),k}x^T_{(i+1),k}\right\rangle + \mathrm{diag}(\langle\psi_{i,k}\rangle)\right)^{-1},$   (19)
         $\mu_{l_{i,k}} = \Sigma_{l_{i,k}}\left(-\langle\upsilon_i\rangle\left\langle x_i x_{(i+1),k}\right\rangle\right),$   (20)
   (d) Compute moment $\langle\omega\rangle$using shaping parameters of (15):
         $\vartheta = \vartheta_0 + \frac{p}{2}, \quad \rho = \rho_0 + \frac{1}{2}\mathrm{tr}\left(\left\langle \mathbf{x}\mathbf{x}^T\right\rangle M^T M\right) - \mathbf{y}^T M\langle\mathbf{x}\rangle + \frac{1}{2}\mathbf{y}^T\mathbf{y},$   (21)
3. Report resulting estimated source term $\langle \mathbf{x}\rangle$
```

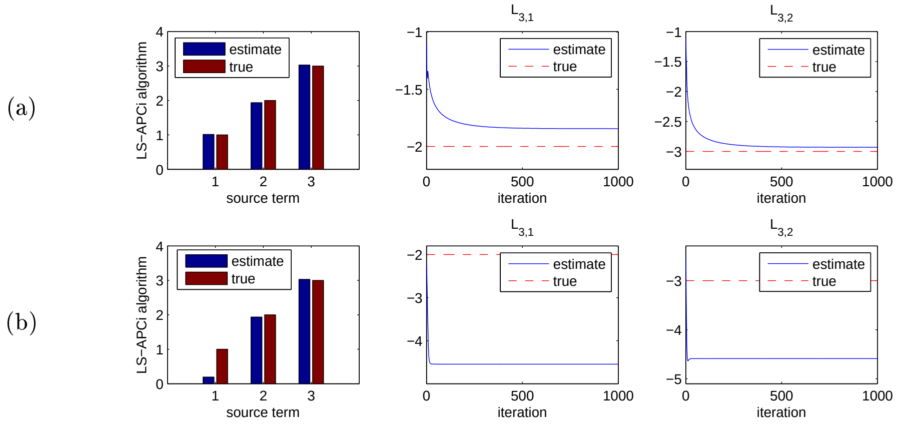

Fig. 2. The results of the LS-APCi algorithm with restricted (a) and unrestricted (b) parameter $\boldsymbol{\upsilon}$.

$\mathbf{x}_{\text{true}} = [1, 2, 3]^T$ and measurement vector is generated according to the assumed model (1) with $\mathbf{e} \sim \mathcal{N}(0, 0.1)$. The negative elements of $\mathbf{y}$ are cropped to 0. We will test two settings of the LS-APCi algorithm: (i) the space of possible solutions is restricted using fixed ratios of elements of vector $\boldsymbol{\upsilon}$: $\boldsymbol{\upsilon} = [\upsilon_1, 10\upsilon_1, 10\upsilon_1]$, and (ii) unrestricted $\boldsymbol{\upsilon}$. The prior intervals for the unknown elements of matrix L are

$$[a_{2,1}, b_{2,1}] = [-10; -1], \quad [a_{3,1}, b_{3,1}] = [-10; -1], \tag{22}$$

while the simulated are $l_{2,1} = -2$ and $l_{3,1} = -3$.

The results of the LS-APCi algorithm are given in Fig. 2. The results suggest that the restriction of the space of possible solutions are beneficial and the estimates converge to the true values, see Fig. 2 (a). On the other hand, estimation of full vector $\boldsymbol{\upsilon} = [\upsilon_1, \upsilon_2, \upsilon_3]$ results in over-parametrization of the problem and the estimates of the ratios in matrix L converge to the centers of the selected intervals. In result, the estimated vector $\mathbf{x}$ differs from the true vector, see Fig. 2 (b).

For comparison, we provide results of the LASSO algorithm, Fig. 3 left, and of the Tikhonov algorithm, Fig. 3 right. Since both algorithms need to preselect suitable regularization parameter, we run both algorithms for a wide range of the regularization parameters and select the best result for each algorithm. The key differences is in estimation of x_1. The LASSO algorithm estimates exact 0 which corresponds to its preference of a sparse solution. The Tikhonov algorithm estimates very similar result to the LS-APCi with unrestricted parameter $\boldsymbol{\upsilon}$. However, the LS-APCi with restriction is clearly closer to the true vector $\mathbf{x}$ as well as to the true covariance matrix and we will use this version of the algorithm in the next experiment.

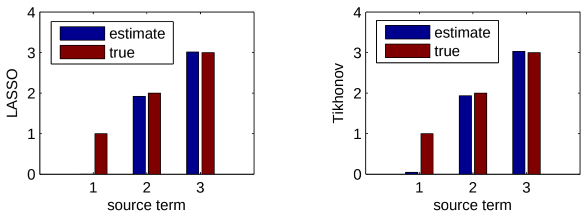

Fig. 3. The results of the LASSO algorithm (left) and Tikhonov algorithm (right).

3.2 Realistic Example

The linear inverse problem (1) is common in estimation of the source term of an atmospheric release. Here, the vector $\mathbf{y}$ contains gamma dose rate (GDR) measurements and the matrix M is a source-receptor-sensitivity matrix computed using an atmospheric transport model [9]. Note that the vector $\mathbf{y}$ does not contain any nuclide-specific information but only sum of GDR of a mixture of nuclides and the matrix M cumulates errors from atmospheric model including errors from the estimates of meteorological conditions (in this case, ECMWF Era-Interim data).

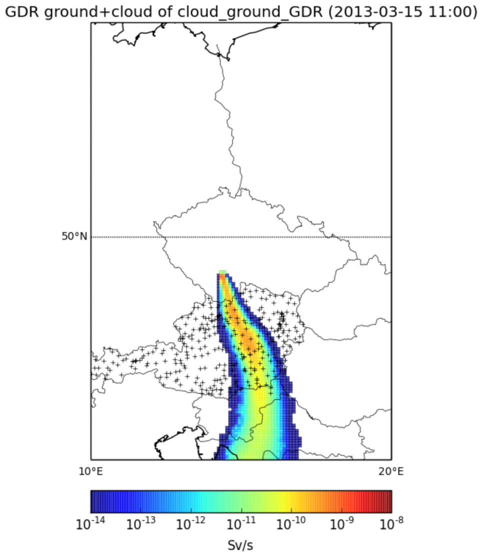

Fig. 4. Gamma dose rate from the cloud shine and deposition.

In this case, a 6 hours long constant rate release is simulated using 3 nuclides: Cs-137, I-131, and Xe-133 from the Czech nuclear power plant Temelin. The Austrian radiation monitoring network is considered to provide measurements from more than 300 receptors implying $M \in R^{4032\times 18}$, see Fig. 4. To simulate realistic conditions, different meteorological data were used for generation matrix M and for generation of simulated measurements $\mathbf{y}$. The problem is critically ill-conditioned and classical optimization methods provide unsuitable results. For our algorithm, we use the following expert-defined intervals of nuclide ratios:

$$[a_{7:12,1}, b_{7:12,1}] = [-10, -3], \quad [a_{13:18,1}, b_{13:18,1}] = [-20, -50], \tag{23}$$

covering the true (simulated) ratios $l_{7:12,1} = -3.8$ and $l_{13:18,1} = -31.3$ (which is, however, unknown in reality).

The results of the LS-APCi algorithm are given in Fig. 5 using subplot for each nuclide. We conclude that the results well correspond to the true releases. Note that in sums of the elements, $\mathbf{x}_{\text{true}}$ and the estimated $\mathbf{x}$ are almost equal. The dissimilarities can be caused by mismatch in the metheorological conditions as well as by uncertainty of the measurement. We perform also run of the LS-APCi algorithm with unrestricted $\boldsymbol{\upsilon}$ with significantly worse results; hence, we conclude that the restriction of $\boldsymbol{\upsilon}$ is crucial for the algorithm.

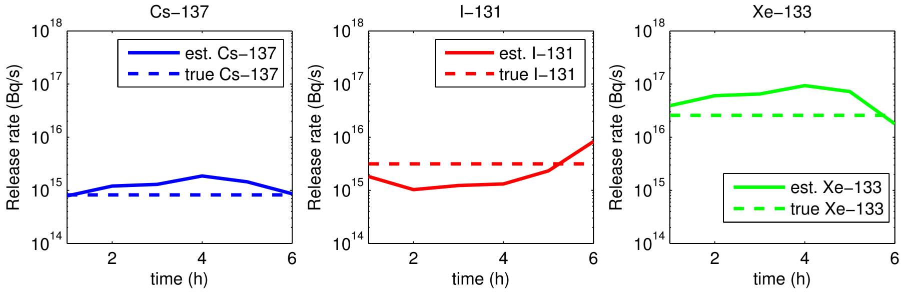

Fig. 5. The results of the source term estimation of 6 hour constant release of 3 nuclides using LS-APCi algorithm.

The results are compared with those of optimization approach with LASSO and Tikhonov regularization with the same ranges restrictions (23) as the LS-APCi algorithm. For this experiment, we used CVX toolbox [4,5] where the optimization problem (2) can be formulated to respect the ranges given in (23). Since the crucial parameter of the optimization approach (2) is α, we run the LASSO and Tikhonov algorithms with $\alpha \in \left[10^{-5}, 10^{5}\right]$. Similarly, we identify as the most significant initial parameter of the LS-APCi algorithm as $\Upsilon = \alpha I_n$; hence, we compare these 3 algorithm with respect to this parameter α. We normalize each nuclide activity to interval $[0, 1]$ and compute mean squared error (MSE) for each α and for each algorithm. The MSE depending on selected parameter α are given in Fig. 6, top, accompanied by the estimated sum of total activity of the source term. From these results, we can identify two main modes of the LS-APCi solution. Note that the natural choice $\Upsilon = I_n$, see Algorithm 1, lies in the correct mode of the solution, see Fig. 5, while the second mode of solution is clearly degenerate. Another situation is in the case of the optimization approaches where continuum of results are observed. Both optimization approaches were able to obtain slightly better results in terms of MSE for specific α; however, it would be difficult to select the correct parameter α without knowledge of the true solution.

4 Conclusion

The linear inverse problem was studied with specific regularization using modeling of a covariance matrix in the modified Cholesky form. We employed the Variational Bayes inference which allows us to deal with vague prior information about range of elements of the covariance matrix using truncated Gaussian prior. We have shown an advantage of the proposed LS-APCi method over the classic optimization approach with LASSO or Tikhonov regularizations. Moreover, we applied the methods to estimation of the source term of atmospheric release from realistic scenario where 6 hours release of mixture of 3 nuclides is simulated. The results suggest that all methods are capable to reach a suitable solution using

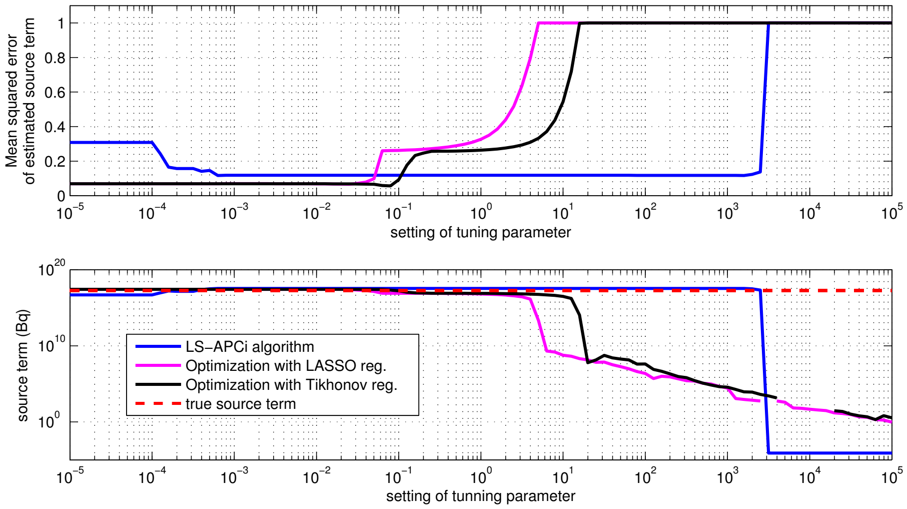

Fig. 6. Top row: mean squared error between the true source term and the estimated source term for each tested algorithm and each parameter α. **Bottom row**: sum of total activity of the source term for each algorithm accompanied by the true sum of the source term (red dashed line).

particular setting of parameters; however, LS-APCi method is much more robust to selection of the tuning parameters.

Acknowledgement

The research leading to these results has received funding from the Norwegian Financial Mechanism 2009-2014 under Project Contract no. MSMT-28477/2014, Source-Term Determination of Radionuclide Releases by Inverse Atmospheric Dispersion Modelling (STRADI). The authors would like to thank to Radek Hofman for data for the realistic experiment.

Appendix

Truncated Gaussian distribution, denoted as $t\mathcal{N}$, of a scalar variable x on interval $[a;b]$ is defined as $t\mathcal{N}_x(\mu,\sigma,[a,b]) = \frac{\sqrt{2}\exp(-\frac{1}{2\sigma}(x-\mu)^2)}{\sqrt{\pi\sigma}(erf(\beta)-erf(\alpha))}\chi_{[a,b]}(x)$, where $\alpha = \frac{a-\mu}{\sqrt{2\sigma}}$, $\beta = \frac{b-\mu}{\sqrt{2\sigma}}$, function $\chi_{[a,b]}(x)$ is a characteristic function of interval $[a,b]$ defined as $\chi_{[a,b]}(x) = 1$ if $x \in [a,b]$ and $\chi_{[a,b]}(x) = 0$ otherwise. erf() is the error function defined as $\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-u^2}\mathrm{d}u$.

The moments of truncated Gaussian distribution are $\langle x\rangle = \mu - \sqrt{\sigma}\frac{\sqrt{2}[\exp(-\beta^2)-\exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta)-\mathrm{erf}(\alpha))}$ and $\langle x^2\rangle = \sigma + \mu\widehat{x} - \sqrt{\sigma}\frac{\sqrt{2}[b\exp(-\beta^2)-a\exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta)-\mathrm{erf}(\alpha))}$. For multivariate case, see [11].

References

1. J.M. Bernardo and A.F.M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.
2. M.J. Daniels and M. Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566, 2002.
3. G.H. Golub, P.C. Hansen, and D.P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.
4. M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.
5. M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, http://cvxr.com/cvx. 2014.
6. K. Khare, B. Rajaratnam, et al. Wishart distributions for decomposable covariance graph models. *The Annals of Statistics*, 39(1):514–555, 2011.
7. M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.
8. O. Saunier, A. Mathieu, D. Didier, M. Tombette, D. Quélo, V. Winiarek, and M. Bocquet. An inverse modeling method to assess the source term of the Fukushima nuclear power plant accident using gamma dose rate observations. *Atmospheric Chemistry and Physics*, 13(22):11403–11421, 2013.
9. P. Seibert and A. Frank. Source-receptor matrix calculation with a lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics*, 4(1):51–63, 2004.
10. V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, 2006.
11. V. Šmídl and O. Tichý. *Sparsity in Bayesian Blind Source Separation and Deconvolution*, volume 8189 LNAI of *Lecture Notes in Computer Science*, pages 548–563. 2013.
12. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
13. O. Tichý and V. Šmídl. Non-parametric bayesian models of response function in dynamic image sequences. *Computer Vision and Image Understanding*, 151:90–100, 2016.
14. M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.