A Bayesian Inverse Solution using Independent Component Analysis

Jouni Puuronen, Aapo Hyvärinen

Dept of Mathematics and Statistics
Dept of Computer Science and HIIT
University of Helsinki, Finland

October 30, 2013
Accepted for publication in Neural Networks

Abstract

We present new results about the simultaneous linear inverse problems using independent component analysis (ICA), which can be used to separate the data into statistically independent components. The idea of using ICA in solving such inverse problems, especially in EEG/MEG context, has been a known topic for at least more than a decade, but the known results have been justified heuristically, and their relationships are not understood properly. Here we show how to obtain a Bayesian posterior for a spatial source distribution, by using an ICA demixing matrix as an input. The posterior enables us to rederive and reinterpret the previously known methods, and also provides completely new methods.

Keywords: Independent component analysis, electroencephalography, magnetoencephalography, Bayesian methods, inverse problem, source localization.

1 Introduction

Our study concerns simultaneous inverse problems that can be formulated in the form

$$X = fZ + \mathcal{E}, \tag{1}$$

which, in sufficiently low frequency range (less than 1 kHz), includes the electro- and magnetoencephalography (EEG/MEG) inverse problems [1] [2]. Here $f \in \mathbb{R}^{n_{\max} \times k_{\max}}$ is a known forward matrix. The number of measurement channels is $n_{\max}$, and the number of points in the spatial source space is $k_{\max}$. We assume $n_{\max} \ll k_{\max}$. The measurement data is $X \in \mathbb{R}^{n_{\max} \times t_{\max}}$, where $t_{\max}$ is the number of time points. These type of inverse problems are often called simultaneous inverse problems to emphasize that $t_{\max} > 1$. The source matrix is $Z \in \mathbb{R}^{k_{\max} \times t_{\max}}$, and an additive noise is $\mathcal{E} \in \mathbb{R}^{n_{\max} \times t_{\max}}$.

We shall assume that we know the distribution of the noise to be Gaussian, each column $\mathcal{E}(t) \in \mathbb{R}^{n_{\max} \times 1}$ being independent with a zero mean and a known covariance Σ. We follow a convention that capital letters denote random variables, and lower case letters denote constant parameters and observed variable values. As a consequence, some vectors are denoted by capital letters, and some matrices by lower case letters, which might seem confusing, but this convention makes the Bayesian approach as clear as possible. [1] When a is a matrix, we denote its rows as a_{i*}, and its columns as a_{*i}. An exception to this is that if the second index is the time parameter, we denote $a(t)$ in place of a_{*t}.

Here we investigate how to use Independent Component Analysis (ICA) to solve these kind of inverse problems. Usually ICA concerns problems of the form

$$X = a\tilde{Z}, \tag{2}$$

where $\tilde{Z} \in \mathbb{R}^{m_{\max} \times t_{\max}}$ contains statistically independent signals on its rows, and $a \in \mathbb{R}^{n_{\max} \times m_{\max}}$ is called the mixing matrix. Here $m_{\max}$ is the assumed number of independent sources. An ICA algorithm takes the observed x as input, and gives as output a demixing matrix $w \in \mathbb{R}^{m_{\max} \times n_{\max}}$ such that the rows of wx are estimates of the rows of z [3] [4]. Our original inverse problem, defined in Equation (1), is not precisely of this form, but is clearly related.

The idea of using ICA as a tool to solve inverse problems of the form (1) has already received wide attention in the EEG/MEG context [5] [4]. One strategy has been to first decompose the measurement data x into a sum of components $x = x^1 + \ldots + x^{m_{\max}} + \varepsilon$, and then use some well-known inverse operator to the components separately [10] [11] [12]. The components are usually obtained by a formula $x^m = (w^+)_{*m} w_{m*} x$, or by some related method. Alternatively the vector $(w^+)_{*m}$ could be used instead of x^m. Here w^+ is a pseudoinverse of the w. Some authors have proposed to use the pseudoinverse w^+ with a philosophy that may not have been obvious in the light of well-known inversion strategies [13]. Also, alternative ways of applying ICA on EEG and MEG, which are not included in the described framework, have been recently introduced [7] [8].

There exists several ways to approach this inverse problem in a Bayesian spirit. One possibility is to attempt to solve an ICA problem with some prior information about the mixing matrix [14] [15] [16] [17]. In this paper we show how to use ordinary ICA as a tool for producing a Bayesian inverse solution. This approach allows us to enhance our understanding of the previously known inverse methods, and the Bayesian posterior also provides a new objective function for the inverse problem.

[1] The noise covariance Σ is an exception as a capital letter without being a random variable, since small sigma would be too confusing in its place.

2 Model and posterior for spatial distribution

2.1 Definition of the model

We start by introducing a probabilistic model for the source matrix Z. We assume that the source Z in (1) can be written as a product

$$Z = SC, \tag{3}$$

where $S \in \mathbb{R}^{k_{\max} \times m_{\max}}$ and $C \in \mathbb{R}^{m_{\max} \times t_{\max}}$ are independent random matrices, with $m_{\max}$ as the assumed number of the independent sources. We assume $m_{\max} \leq n_{\max}$. Now the X defined in Equation (1) becomes

$$X = fSC + \mathcal{E}. \tag{4}$$

The notation is motivated by the words "spatial" for S and "chronos" for C. A column of Z, denoted by $Z(t)$, can be written as

$$Z(t) = \sum_{m=1}^{m_{\max}} S_{*m} C_m(t) \ \in \ \mathbb{R}^{k_{\max} \times 1}. \tag{5}$$

Now the columns $S_{*m} \in \mathbb{R}^{k_{\max} \times 1}$ describe the spatial distributions of individual sources, and the rows $C_m \in \mathbb{R}^{1 \times t_{\max}}$ describe the time courses. The spatial part S will obey some prior $p(S = s)$, which can be specified later. Below we will consider several possible spatial priors. The temporal part C will obey a prior such that its rows are statistically independent, and they also possess some property, such as non-Gaussianity, that can be exploited for the purpose of blind source separation, which here means estimation of the ICA model. We do not need to specify what that property is, since below we simply assume that such a blind source separation is possible and has been done.

The assumption (3) makes sense, since in many inverse problems the original sources have in a some sense fixed spatial patterns despite the time dependence. Also we usually assume Z to be very sparse with respect to the spatial parameter k, but only moderately sparse, or otherwise non-Gaussian, with respect to the time parameter t. Hence we assume Z to be a sum of terms $S_{*m}C_m$, where S_{*m} are very sparse, and C_m only moderately sparse or otherwise non-Gaussian.

We also assume that C is white, for the purpose of reducing the ambiguities in S and C. The product SC will still be invariant under a transformation that permutates the columns of S and rows of C similarly, and also under the change of signs of the columns and rows. These invariances do not imply serious difficulties.

Our approach begins with the assumption that we have already obtained a demixing matrix $w \in \mathbb{R}^{m_{\max} \times n_{\max}}$ by applying some ICA or blind source separation algorithm on the data to estimate the model (2), and that we have also guessed correctly the number of independent sources $m_{\max}$. We can then assume that $C = wx$ holds, and consider the Bayesian posterior for S. The idea behind this assumption is that if the noise is small so that $x \approx fsc$, as implied by (4), the demixing matrix w will behave as a left inverse of the fs, and $wx \approx c$ will hold. By using the definition of a conditional probability, the formula $X = fSC + \mathcal{E}$, and the assumption that S, C and $\mathcal{E}$ are independent, we obtain

$$\begin{aligned} p(S = s \mid X = x,\ C = wx) &= \frac{p(S = s,\ fSC + \mathcal{E} = x,\ C = wx)}{p(X = x,\ C = wx)} \\ &= \frac{p(S = s)p(C = wx)}{p(X = x, C = wx)} p\big(\mathcal{E} = (\mathrm{id} - fsw)x\big) \end{aligned} \tag{6}$$

Here id is an $n_{\max} \times n_{\max}$ identity matrix. By using the assumed distribution for noise, and by ignoring all terms not depending on s, we obtain the logarithmic posterior for S.

$$\begin{aligned} &\log\ p(S = s \mid x, w) \\ &= \log\ p(S = s)\ -\ \frac{1}{2} \sum_{t=1}^{t_{\max}} x(t)^T (\mathrm{id} - fsw)^T \Sigma^{-1} (\mathrm{id} - fsw) x(t)\ +\ \text{const.} \end{aligned} \tag{7}$$

The logarithmic likelihood for s can be obtained from Equation (7) by omitting the prior term $\log\ p(S = s)$ and the constants, and we denote the logarithmic likehood as $\ell(s|w, x, \Sigma)$. The logarithmic likelihood can be written in several different ways, for example:

$$\begin{aligned} \ell(s|w, x, \Sigma) &= -\frac{1}{2} \mathrm{Tr}\Big(xx^T (\mathrm{id} - fsw)^T \Sigma^{-1} (\mathrm{id} - fsw) \Big) \\ &= -\frac{1}{2} \sum_{n,n'=1}^{n_{\max}} \big((\mathrm{id} - fsw) xx^T \big)_{n'n} \big(\Sigma^{-1} (\mathrm{id} - fsw) \big)_{n'n} \end{aligned} \tag{8}$$

We shall compare this likelihood with a simpler objective function $\mathcal{L}_{\mathrm{old}}$, which we define as

$$\mathcal{L}_{\mathrm{old}}(s|w^+) = -\frac{1}{2} \sum_{m=1}^{m_{\max}} \| fs_{*m} - (w^+)_{*m} \|^2. \tag{9}$$

Here w^+ is a pseudoinverse of w. The Moore-Penrose pseudoinverse would be a natural choice, but below we will consider alternatives too. The idea of this simpler objective function is similar to the ideas already studied by

many authors [10] [11] [12]. We shall not attempt to reproduce precisely the same methods as explained by these earlier authors, but instead we compare our new objective function to $\mathcal{L}_{\text{old}}$, which summarises these earlier methods compactly.

2.2 Theoretical analysis

There are several remarks which can be made about the properties of ℓ and $\mathcal{L}_{\text{old}}$. We begin with the simplest ones.

Basic motivation for $\mathcal{L}_{\text{old}}$ In Equation (2) we recalled the ordinary ICA model $X = a\tilde{Z}$. Usually the mixing and demixing matrices a and w are related in such way that they are some pseudoinverses of each other. For example, we could define an estimate of the mixing matrix by a formula $\hat{a} = w^+$.

In our model $X = fSC + \mathcal{E}$ a matrix fs takes the role of a, so we should expect $fs \approx w^+$. This is probably the simplest way to justify $\mathcal{L}_{\text{old}}$.

If $m_{\max} < n_{\max}$ holds, w will not be a square matrix, and it will have several different pseudoinverses. This introduces the problem of choosing the most optimal one.

If the noise is very small, the Moore-Penrose pseudoinverse will be the most natural choice. This can be justified by noticing that ICA algorithms usually produce such w that $w \operatorname{Im}(a)^\perp = \{0\}$ holds. Here $\operatorname{Im}(a)^\perp$ is the orthogonal complement of the image of a. This follows from the PCA dimension reduction, and implies that w is actually the Moore-Penrose pseudoinverse of a. Then $w = a^+$ implies $w^+ = a$.

It turns out that the Moore-Penrose pseudoinverse is not necessarily the best estimate for the mixing matrix when non-trivial noise is present. We will discuss this in more detail in Section 4.5.

Relation in the limit of small noise Another simple remark is that if the noise is extremely small, an approximation $x \approx w^+ wx$ will hold when w^+ is defined as the Moore-Penrose pseudoinverse. This is because $w^+ w$ is roughly the orthogonal projection matrix to the $m_{\max}$-dimensional subspace where the relevant input data resides (while ww^+ is the $m_{\max} \times m_{\max}$ identity matrix). This implies the following approximation.

$$(\mathrm{id} - fsw)x \;\approx\; (w^+ - fs)wx \tag{10}$$

Thus we see that if $\mathcal{L}_{\text{old}}$ is close to zero, ℓ must be close to zero too. This implication cannot be justified so easily if the noise is not small.

Maximization of $\mathcal{L}_{\text{old}}$ Since we assume $n_{\max} < k_{\max}$, for simplicity we can also assume that the rows of f are linearly independent. This implies that $\mathcal{L}_{\text{old}}$ reaches its maximal value, which is zero, at the estimate $\hat{s} = f^+ w^+$. In this formula f^+ can be any right inverse of f, while w^+ must be the same pseudoinverse which is present in (9).

In applications such as MEG/EEG source localization and other inverse problems the rows of f can be strongly linearly correlated, and the pseudoinverse f^+ will need to be regularized to be useful. For simplicity, in our theoretical analysis we assume that the precise pseudoinverses would work. In some special cases the rows could be precisely linearly dependent too, but we omit this case from our analysis.

Continuing with the assumption of linear independence, the set in which $\mathcal{L}_{\text{old}}$ is maximized is

$$f^+ w^+ + \ker(f)^{m_{\max}}. \tag{11}$$

Here $\ker(f)^{m_{\max}}$ means the set of $k_{\max} \times m_{\max}$ matrices whose columns are in $\ker(f) \subset \mathbb{R}^{k_{\max} \times 1}$. This can be proven easily by writing a point of maximum in a form $s_{*m} = f^+(w^+)_{*m} + \Delta s_{*m}$. It follows that

$$\mathcal{L}_{\text{old}}(s|w^+) = -\frac{1}{2} \sum_{m=1}^{m_{\max}} \|f \Delta s_{*m}\|^2, \tag{12}$$

and we see that the Δs_{*m} must be contained in the kernel of f.

Limited effect of assumed noise covariance Some information about the likelihood can be obtained by examining its gradient, which is

$$\nabla_s \ell(s|w, x, \Sigma) = f^T \Sigma^{-1} (\text{id} - fsw) x x^T w^T. \tag{13}$$

A surprising remark can be made at this point. If we assume that the rows of f are linearly independent, and that Σ is finite and non-singular, the input parameter Σ actually has no effect on the maxima of ℓ. This follows from the fact that $f^T \Sigma^{-1}$ has no non-trivial kernel, and $\nabla_s \ell = 0$ is equivalent with a relation $(\text{id} - fsw) x x^T w^T = 0$. Notice that the parameter x will still have an effect on the maxima, and on the other hand a relation $\frac{1}{t_{\max}} x x^T \approx f s s^T f^T + \Sigma^*$ will hold, where the Σ^* is the true noise covariance. Thus, the true noise covariance will have an effect on the maxima, regardless of the model parameter Σ. Also, if some of the eigenvalues of model parameter Σ are close to zero, numerical effects on the estimated maxima can occur.

Maximization of likelihood Next we generalize the estimate $\hat{s} = f^+ w^+$.

Theorem 1. *Assume that the rows of f and wx are linearly independent (and that Σ is finite and non-singular). The estimate $\hat{s}^0$ defined by a formula*

$$\hat{s}^0 = f^+ x x^T w^T (w x x^T w^T)^{-1} \tag{14}$$

maximizes the likelihood $\ell(s|w,x,\Sigma)$. Here f^+ can be any right inverse of f. The set in which ℓ is maximized is

$$\hat{s}^0 + ker(f)^{m_{max}}. \tag{15}$$

Proof is given later in Section 3.

Notice that if $w^+wx \approx x$ holds, we can substitute w^+w between f^+ and x in Equation (14), and the formula simplifies back to $\hat{s}^0 \approx f^+w^+$. We see that (14) generalizes the estimate f^+w^+ by taking into account the noise in a non-trivial way. In fact, we have even a stronger result concerning the estimate f^+w^+:

Theorem 2. *Assume that the rows of f and w are linearly independent. Also assume that the sample covariances are precisely the theoretical values, so that $cc^T = t_{max}id$, $\varepsilon\varepsilon^T = t_{max}\Sigma$, and $c\varepsilon^T = 0$. If the covariance of noise is proportional to the identity, so that $\Sigma = \sigma^2 id$ with some real $\sigma^2 > 0$, the likelihood will reach its maximal value at $\hat{s} = f^+w^+$. Here f^+ can be any right inverse of f, while w^+ must be the Moore-Penrose pseudoinverse of w.*

Above we noted that if the noise is very small, the estimate $\hat{s} = f^+w^+$ will maximize the likelihood. Now Theorem 2 states that actually the noise does not need to be very small, but instead it is sufficient that the noise covariance is proportional to the identity.

The assumption that the sample covariances are precisely the theoretical values is slightly hypothetical, but it is a reasonable approximation that if $t_{\max}$ is sufficiently large, and if the sample assumption holds approximately to a sufficent degree, the estimate $\hat{s} = f^+w^+$ will be approximately a point of maximum. Empirical simulations support this hypothesis.

In the light of what we now know, the condition $\Sigma = \sigma^2$id (with the true noise covariance, not only the input parameter) implies the maxima of ℓ and $\mathcal{L}_{\text{old}}$ to be the same (assuming we are using the Moore-Penrose pseudoinverse as w^+). In other words, it makes sense to bother with the likelihood only when non-trivial noise is present.

The concavity of an objective function is an important property if we are interested in its maximization. To this end, we have the following result:

Theorem 3. *The log-likelihood $s \mapsto \ell(s|w,x,\Sigma)$ is concave, meaning that for all $\tilde{s}, \overline{s} \in \mathbb{R}^{k_{max} \times m_{max}}$ and $0 \leq \alpha \leq 1$ inequality*

$$\ell(\alpha\tilde{s} + (1-\alpha)\overline{s}|w,x,\Sigma) \;\geq\; \alpha\ell(\tilde{s}|w,x,\Sigma) \;+\; (1-\alpha)\ell(\overline{s}|w,x,\Sigma) \tag{16}$$

holds.

Since a sum of two concave functions is also a concave one, we see that the logarithmic posterior of s will always be concave when the logarithmic prior is chosen concave first. In fact also the $\mathcal{L}_{\text{old}}(s|w^+)$ is concave, but this is not a novel result.

3 Proofs of theorems

3.1 Proof of Theorem 1

Theorem 1 can be proven very mechanically by examining the gradient given in Equation (13) once it has been first proven that the log-likelihood is maximized globally where the gradient is zero. The second partial derivatives of the log-likehood have the following formula:

$$\frac{\partial^2 \ell}{\partial s_{k'm'} \partial s_{km}} = -(f^T \Sigma^{-1} f)_{kk'} (wxx^T w^T)_{m'm} \tag{17}$$

We see that the second partial derivatives are constants with respect to the s, and the log-likelihood must be some quadratic form. Let us identify s, which is a $k_{\max} \times m_{\max}$ matrix, with a $k_{\max} m_{\max} \times 1$ vertical vector, and denote the $k_{\max} m_{\max} \times k_{\max} m_{\max}$ Hessian matrix as $\nabla^2 \ell$. Now with a brief calculation we get

$$\begin{aligned} s^T (\nabla^2 \ell) s &= \sum_{k,k'=1}^{k_{\max}} \sum_{m,m'=1}^{m_{\max}} s_{k'm'} \frac{\partial^2 \ell}{\partial s_{k'm'} \partial s_{km}} s_{km} \\ &= -\operatorname{Tr}\big((fswx)^T \Sigma^{-1} fswx\big) \leq 0. \end{aligned} \tag{18}$$

We see that the eigenvalues of the Hessian matrix must be either negative or zero, and the log-likelihood is a paraboloid opening downwards (at least non-properly).

We can now proceed by examining the zeros of the gradient.

$$\begin{aligned} \nabla_s \ell(s|w,x,\Sigma) = 0 \quad &\iff \quad f^T \Sigma^{-1} (\mathrm{id} - fsw) xx^T w^T = 0 \\ &\iff \quad (\mathrm{id} - fsw) xx^T w^T = 0 \\ &\iff \quad xx^T w^T = fswxx^T w^T \\ &\iff \quad xx^T w^T (wxx^T w^T)^{-1} = fs \\ &\impliedby \quad f^+ xx^T w^T (wxx^T w^T)^{-1} = s \end{aligned} \tag{19}$$

Here we used the fact that $f^T \Sigma^{-1}$ has no non-trivial kernel, and the fact that $wxx^T w^T$ is invertible, which follow from the assumptions. Matrices like $wx(wx)^T$ are always diagonalizable with real eigenvalues by symmetry, and can never have negative eigenvalues. The assumption that the rows of wx are linearly independent implies that $wx(wx)^T$ cannot have zero eigenvalues either, so the matrix must be invertible.

If $\tilde{s}$ is some maximum distinct from $\hat{s}^0$, we can proceed from the formulas in (19) as follows.

$$xx^T w^T (wxx^T w^T)^{-1} = f\tilde{s} \quad \implies \quad \hat{s}^0 = f^+ f \tilde{s} \tag{20}$$

Here we multiplied the both sides from left by f^+, and used the definition of $\hat{s}^0$. On the other hand, the definition of $\hat{s}^0$ implies $f^+ f\hat{s}^0 = \hat{s}^0$. So we get

$$f^+ f\hat{s}^0 = f^+ f\tilde{s} \quad \Longrightarrow \quad f(\hat{s}^0 - \tilde{s}) = 0. \tag{21}$$

Here we used the fact that f^+ has no non-trivial kernel, and obtained the desired result $\tilde{s} \in \hat{s}^0 + \ker(f)^{m_{\max}}$.

3.2 Proof of Theorem 2

We begin with the lower formula of Equation (8), and substitute

$$xx^T = t_{\max}(fss^T f^T + \Sigma), \tag{22}$$

which follows from the assumption that sample covariances are precisely the theoretical values. We obtain

$$\begin{aligned}\ell(s|w,x,\Sigma) = -\frac{t_{\max}}{2}\Big(&\underbrace{\sum_{n,n'=1}^{n_{\max}} \big((\mathrm{id} - fsw)fss^T f^T\big)_{n'n}\big(\Sigma^{-1}(\mathrm{id} - fsw)\big)_{n'n}}_{\text{1st term}} \\ &+ \underbrace{\sum_{n,n'=1}^{n_{\max}} \big((\mathrm{id} - fsw)\Sigma\big)_{n'n}\big(\Sigma^{-1}(\mathrm{id} - fsw)\big)_{n'n}}_{\text{2nd term}}\Big).\end{aligned} \tag{23}$$

Next, we shall prove that both of these two terms separately reach minimal values at $s = f^+ w^+$. It turns out that for the 1st term, the assumption about Σ will not be needed. Firstly, it is a simple exercise to check that the 1st term is zero when $s = f^+ w^+$. One only needs to use $ww^+ = \mathrm{id}$ and $ff^+ = \mathrm{id}$ a few times after substitution. On the other hand, the 1st term can be manipulated into the form

$$\mathrm{Tr}\big(s^T f^T (\mathrm{id} - fsw)^T \Sigma^{-1}(\mathrm{id} - fsw)fs\big). \tag{24}$$

Since Σ^{-1} is symmetric with non-negative eigenvalues, the 1st term can never reach negative values. So we can deduce that the 1st term reaches its minimal value at $s = f^+ w^+$.

When proving that the 2nd term reaches minimal value at $s = f^+ w^+$, we must first use the assumption $\Sigma = \sigma^2 \mathrm{id}$. The 2nd term then becomes

$$\sum_{n,n'=1}^{n_{\max}} (\mathrm{id} - fsw)^2_{n'n}. \tag{25}$$

At this point it is a good idea to examine a function

$$J(g) = \sum_{n,n'=1}^{n_{\max}} (\mathrm{id} - g)^2_{n'n}, \tag{26}$$

where id and g are $n_{\max} \times n_{\max}$ matrices. Suppose we want to minimize $J(g)$ with a constraint $\dim(\mathrm{Im}(g)) = m_{\max}$. The result will turn out to be that if g is an orthogonal projection to some $m_{\max}$-dimensional subspace, the $J(g)$ obtains its minimal value, which turns out to be $n_{\max} - m_{\max}$. First, the easy part is to check that if g is an orthogonal projection to some $m_{\max}$-dimensional subspace, then $J(g) = n_{\max} - m_{\max}$. This can be proven by using the fact that in this case g must be symmetric, and it can be diagonalized with an orthogonal transformation. In the definition of $J(g)$ we might as well replace the g with its diagonalized form. On the other hand g's eigenvalues must include $m_{\max}$ "1":s and $n_{\max} - m_{\max}$ "0":s.

The more difficult part of the proof is to show that even when nothing else but $\dim(\mathrm{Im}(g)) = m_{\max}$ is assumed, still $J(g)$ cannot reach values smaller than $n_{\max} - m_{\max}$. In this case, assume that $g = u\lambda v^T$ is the singular value decomposition of g. In the standard form, all matrices u, λ, v would be $n_{\max} \times n_{\max}$. However, λ must have $n_{\max} - m_{\max}$ "0":s on its diagonal, by the assumption $\dim(\mathrm{Im}(g)) = m_{\max}$. Therefore, we can omit the redundant rows and columns of these matrices, and replace them with $n_{\max} \times m_{\max}$ matrices u and v, and $m_{\max} \times m_{\max}$ diagonal matrix λ. The identities $u^T u = \mathrm{id}$ and $v^T v = \mathrm{id}$ still hold. After some work, a following formula can be obtained:

$$\begin{aligned} J(g) &= \sum_{n,n'=1}^{n_{\max}} \Big(\delta_{nn'} - \sum_{m=1}^{m_{\max}} u_{nm}\lambda_{mm}(v^T)_{mn'}\Big)^2 = \cdots \\ \cdots &= n_{\max} + \sum_{m=1}^{m_{\max}} \Big(\lambda_{mm}^2 - 2\big((v^T)_{m*}u_{*m}\big)\lambda_{mm}\Big) \end{aligned} \tag{27}$$

Since there are no constraints between λ and (u, v), an equation $\frac{\partial}{\partial \lambda_{mm}} J = 0$ must hold at the point of minimum. When computing the partial derivative, $(v^T)_{m*}u_{*m}$ can be considered as a real constant. We find that

$$\lambda_{mm} = (v^T)_{m*}u_{*m} \tag{28}$$

must hold at the point of minimum, which in turn implies

$$J(g) = n_{\max} - \sum_{m=1}^{m_{\max}} \lambda_{mm}^2. \tag{29}$$

Finally, the Cauchy-Schwarz inequality $|\lambda_{mm}| \leq \|v_{*m}\|\|u_{*m}\| = 1$ confirms that $J(g)$ cannot obtain values smaller than $n_{\max} - m_{\max}$.

Now we are ready to deal with Equation (25). The product fsw is an $n_{\max} \times n_{\max}$ matrix which always fulfills the relation $\dim(\mathrm{Im}(fsw)) \leq m_{\max}$, since fs is $n_{\max} \times m_{\max}$ and w is $m_{\max} \times n_{\max}$. The assumption $s = f^+w^+$ implies $fsw = w^+w$, and w^+w actually is an orthogonal projection to an $m_{\max}$-dimensional subspace, so the quantity in Equation (25) must obtain the minimal value at $s = f^+w^+$.

3.3 Proof of Theorem 3

A function, whose graph is a paraboloid opening downwards (properly or non-properly) is always concave. The log-likelihood is defined in such way, that it is not manifestly a quadratic form with respect to the s, since the matrix s is between matrices f and w in a way which has no obvious intuitive meaning. However, it can be proven that the log-likelihood is indeed a quadratic form with respect to the s, and also that the graph is a paraboloid opening downwards (non-properly). In fact, this was done in the proof of Theorem 1.

So the easiest way to prove Theorem 3 is to make the remark that in the proof of Theorem 1 we already proved the log-likelihood to be a quadratic form with a Hessian matrix that has no positive eigenvalues.

It is also possible to prove Theorem 3 more mechanically, without using the proof of Theorem 1, by using the inequality

$$|a^T\Sigma^{-1}b| \leq \sqrt{a^T\Sigma^{-1}a}\sqrt{b^T\Sigma^{-1}b}, \tag{30}$$

which holds with arbitrary vectors $a, b \in \mathbb{R}^{n_{\max}}$. This is simply the Cauchy-Schwarz inequality with an inner product defined by Σ^{-1}. We omit the details of this proof, since this way is not very attractive after the previous one.

4 Spatial estimate methods

Next, we consider practical methods ensuing from the model and theory of the preceding sections.

4.1 Sparse prior with gradient ascent

One of the most obvious ways to maximize the logarithmic posterior is the gradient ascent, where we allow the quantity s to evolve by steps

$$s(i+1) = s(i) + \mu\nabla_s \log\ p\big(s(i)|x, w\big). \tag{31}$$

The origin $s(0) = 0$ is one good starting point. In order to use the gradient ascent, the prior $p(S = s)$ must be specified. A well-known approach is to

use a prior defined as

$$\log\ p(S=s) = -\lambda \|s\|_{1,\varepsilon_{\mathrm{reg}}} = -\lambda \sum_{m=1}^{m_{\max}} \sum_{k=1}^{k_{\max}} \sqrt{s_{km}^2 + \varepsilon_{\mathrm{reg}}^2} \tag{32}$$

[18]. This is roughly the same as $\log\ p(S=s) = -\lambda\|s\|_1$, which is a popular sparsity favouring prior, but we have arranged it differentiable at the origin with a regularizing parameter $\varepsilon_{\mathrm{reg}} > 0$. The partial derivatives of the prior are

$$\frac{\partial}{\partial s_{km}} \log\ p(S=s) = -\frac{\lambda s_{km}}{\sqrt{s_{km}^2 + \varepsilon_{\mathrm{reg}}^2}}. \tag{33}$$

The purpose of the parameter λ is to allow us to adjust the strength of the prior.

The gradient of the logarithmic likelihood was given in Equation (13). The objective function $\mathcal{L}_{\mathrm{old}}$ defined in (9) can also be maximized by a gradient ascent, since its gradient is given by a simple formula

$$\nabla_s \mathcal{L}_{\mathrm{old}}(s|w^+) = f^T(w^+ - fs). \tag{34}$$

So both of the objective functions

$$\ell(s|w,x,\Sigma) - \lambda\|s\|_{1,\varepsilon_{\mathrm{reg}}} \qquad \text{and} \qquad \mathcal{L}_{\mathrm{old}}(s|w^+) - \lambda\|s\|_{1,\varepsilon_{\mathrm{reg}}} \tag{35}$$

can be simply maximized by a gradient ascent.

4.2 Prior with an infinitesimal coefficient

One problem with the gradient ascent maximization of the posterior is that there is no obvious way to decide a value for the prior coefficient λ. One mathematically consistent estimate can be defined by defining λ as some positive infinitesimal. This is equivalent to demanding that ℓ or $\mathcal{L}_{\mathrm{old}}$ is maximized first, and we then mimimize $\|s\|_{1,\varepsilon_{\mathrm{reg}}}$ while maintaining the constraint that ℓ or $\mathcal{L}_{\mathrm{old}}$ is kept at the maximal value.

It would be a mistake to search for this estimate by the ordinary gradient ascent with some very small λ, since such approach would turn out to be very slow. The reason for this is that the maximal rate of convergence of an ordinary gradient ascent is bounded by the ratio of the largest and the smallest eigenvalues of the Hessian matrix at the point of maximum. We know that some of the eigenvalues of the Hessian matrices of ℓ and $\mathcal{L}_{\mathrm{old}}$ are precisely zero. The Hessian matrices at the maxima of the functions in (35) will not have zero eigenvalues when $\lambda > 0$, but in the limit $\lambda \to 0$ some eigenvalues will approach zero.

On the other hand, we have proven earlier that the sets where ℓ and $\mathcal{L}_{\mathrm{old}}$ are maximized are of the form $\hat{s}^0 + \ker(f)^{m_{\max}}$, and we have also found

analytical formulas for $\hat{s}^0$. The formula $\hat{s}^0 = f^+ w^+$ provides a maximum for $\mathcal{L}_{\text{old}}$, and this formula was generalized in Equation (14) to give a maximum for ℓ. Thus, if we denote as $P_{\ker(f)}$ the orthogonal projection to the kernel of f, we can set up a following iterative method: First set $s(0) = \hat{s}^0$, and then use the recursive formula

$$s(i+1) = s(i) - \mu P_{\ker(f)} \nabla_s \|s(i)\|_{1,\varepsilon_{\text{reg}}}. \tag{36}$$

The pseudoinverses and projection matrices may have to be regularized before numerical use. We propose using some integer parameter n_{reg} such that $1 \leq n_{\text{reg}} \leq n_{\max}$, and defining $f^+_{n_{\text{reg}}}$ by the formula $f^+_{n_{\text{reg}}} = v \Lambda^{-1}_{n_{\text{reg}}} u^T$, where $f = u \Lambda v^T$ is the singular value decomposition, and $\Lambda^{-1}_{n_{\text{reg}}}$ has been defined by setting $(\Lambda^{-1}_{n_{\text{reg}}})_{nn} = \frac{1}{\Lambda_{nn}}$ for the n_{reg} largest singular values, and $(\Lambda^{-1}_{n_{\text{reg}}})_{nn} = 0$ for other diagonal entries. We can define $v_{n_{\text{reg}}}$ as a $k_{\max} \times n_{\text{reg}}$ matrix, where the columns are those columns of v which correspond to the n_{reg} largest singular values. Then the projection matrix $P_{\ker(f)}$ is naturally regularized by a formula $P_{n_{\text{reg}}} = \text{id} - v_{n_{\text{reg}}} v^T_{n_{\text{reg}}}$.

We do not recommend regularizing the pseudoinverse by some formula such as $(\Lambda^{-1}_{\delta})_{nn} = \Lambda_{nn}/(\Lambda^2_{nn} + \delta^2)$, because there would be no obvious way to regularize the projection $P_{\ker(f)}$ in a compatible manner.

4.3 The method of translating point sources

One possible alternative to the gradient ascent is to make the prior assumption that all sources are strictly point sources. The matrix s can then be parametrized by $2m_{\max}$ numbers $k_1, \ldots, k_{m_{\max}}, a_1, \ldots, a_{m_{\max}}$, so that $s_{km} = a_m \delta_{k,k_m}$. Here $k_m \in \{1, 2, \ldots, k_{\max}\}$ are the locations of the point sources, and $a_m \in \mathbb{R}$ are the amplitudes. Parametrizing s like this is equivalent to a prior $p(S = s)$ being such that it requires s to be of this parametrized form. We can then carry out a local search procedure, where the algorithm simply checks if the likelihood can be increased by making small changes to the parameters, and by flipping the signs of the amplitudes a_m. This method requires that we know which points in the set $\{1, 2, \ldots, k_{\max}\}$ are considered neighbour points, and lacks some generality, but we have found this approach to work very well in simulations where the real source distributions actually were point sources.

In some sense the method described here is more primitive than some previously published methods, which have a similar setting [19]. This primitive approach was sufficient for our purposes, which are explained later in Section 6.

Both the gradient ascent, with real or infinitesimal λ, and the method of spatially translating point sources, can be interpreted as methods to maximize the posterior (7), but with different priors $p(S = s)$.

4.4 The gradient of the likelihood at the origin

Next we compare our work to that of Hild and Nagarajan (H&N) [13], and develop a natural generalization to their method. In a simplified form, the H&N method is to first define a mapping

$$k \mapsto |(w^+)^T_{m*} \overline{f}_{*k}|, \tag{37}$$

and then to examine where it obtains large values, for example by defining an estimate $\hat{k}_m$ as the point were this mapping is maximized. [2] If the true source component s_{*m} is very point like, the $\hat{k}_m$ will usually be very close to the true source position. Here the matrix $\overline{f}$ has been defined by a formula $\overline{f}_{*k} = \frac{f_{*k}}{\|f_{*k}\|_2}$. We found in simulations that the matrix $\overline{f}$ usually works significantly better than the original f. Hild and Nagarajan used this normalization too.

Equation (34) implies

$$\nabla_s \mathcal{L}_{\text{old}}(0|w^+) = f^T w^+. \tag{38}$$

Thus we see that if we replace f by $\overline{f}$, the simplified H&N method is equivalent to simply considering the gradient of the $\mathcal{L}_{\text{old}}$ at the origin $s = 0$, and examining where it is extremized. This observation paves a way for an obvious generalization, since we can exploit the information in x and Σ by considering the gradient of the $\ell(s|w, x, \Sigma)$ at the origin $s = 0$ in a similar manner. The formula is

$$\nabla_s \ell(0|w, x, \Sigma) = f^T \Sigma^{-1} x x^T w^T. \tag{39}$$

The relation of (39) to the H&N method becomes clearer when we prove that if the noise is very small, an equation

$$\frac{1}{t_{\max}} x x^T w^T \approx w^+ \tag{40}$$

will hold when w^+ is the Moore-Penrose pseudoinverse. In order to understand this, let us assume that an equation $x = a\tilde{z}$ holds with some mixing matrix a, and that $\frac{1}{t_{\max}} \tilde{z}\tilde{z}^T = \text{id}$. Then $xx^T = t_{\max} a a^T$ holds too, and if w is a left inverse of a, we get

$$\frac{1}{t_{\max}} w x x^T = \frac{1}{t_{\max}} w (t_{\max} a a^T) = a^T. \tag{41}$$

[2] In their original paper, Hild and Nagarajan divided the spatial source space into blocks of three points, like $\{\{1, 2, 3\}, \{4, 5, 6\}, \ldots\}$, and then carried out constrained maximization procedures in these blocks. These details were relevant for their final algorithm, but not for our discussion. It should be obvious that all methods which we discuss here, can be modified and rigged later.

We explained earlier in Section 2.2 that usually $w^+ = a$ holds (at least, if w has been obtained from a usual ICA algorithm), so we see that $\frac{1}{t_{\max}} wxx^T = (w^+)^T$ holds usually too. We have now proven the approximation (40), and also the result, that if $\Sigma = \sigma^2 \mathrm{id}$ with some very small σ^2, the gradients $\nabla_s \mathcal{L}_{\mathrm{old}}(0|w^+)$ and $\nabla_s \ell(0|w, x, \Sigma)$ differ only by some multiplicative constant.

We now propose the mapping

$$k \mapsto |(wxx^T \Sigma^{-1} \overline{f})_{mk}| \tag{42}$$

as a generalization to the old mapping shown in (37). Using the gradient of the log-likelihood introduces a modification of the Hild-Nagarajan method, which again uses the information in the noise covariance matrix unlike the original method.

4.5 Pseudoinverse of the demixing matrix

Many of our formulas involve a $n_{\max} \times m_{\max}$ matrix w^+, which we call a pseudoinverse of w. There exists two ways to compute this as a Moore-Penrose pseudoinverse, and we now emphasize some important technical details related to this.

A demixing matrix w, whose size is $m_{\max} \times n_{\max}$, can be thought to have been defined by a formula $w = w_{\mathrm{select}} w_{\mathrm{demix}} w_{\mathrm{reduce}}$. Here w_{reduce} is a $d_{\max} \times n_{\max}$ matrix which has been obtained by PCA, and the coefficient $d_{\max}$ denotes the dimension of the subspace where the significant amount of the variance resides. Always $m_{\max} \leq d_{\max} \leq n_{\max}$. The matrix w_{demix} is a $d_{\max} \times d_{\max}$ square matrix which contains whitening and the actual demixing. The $d_{\max} \times n_{\max}$ matrix $w_{\mathrm{demix}} w_{\mathrm{reduce}}$ contains $d_{\max}$ rows of which $m_{\max}$ are relevant, and actually demix the independent components, while $d_{\max} - m_{\max}$ are irrelevant, and only demix noise. So finally the selection matrix w_{select} is defined as a $m_{\max} \times d_{\max}$ matrix with all other entries zero, but $m_{\max}$ "1"-elements at right places so that the irrelevant components are omitted.

One way to compute w^+ is to simply compute the Moore-Penrose pseudoinverse of the w. A second way is to first compute the Moore-Penrose pseudoinverse $(w_{\mathrm{demix}} w_{\mathrm{reduce}})^+$, which will be a $n_{\max} \times d_{\max}$ matrix, and arrive at the $n_{\max} \times m_{\max}$ matrix by as a last step omitting the irrelevant columns.

It turns out that in general the procedures of omitting rows or columns, and of computing the Moore-Penrose pseudoinverse, do not commute. We verified in our simulations that the results can turn out to be inferior if the rows are omitted first, and the Moore-Penrose pseudoinverse computed last. We now emphasize that if $m_{\max} < d_{\max}$ holds, it will be very important to compute the Moore-Penrose pseudoinverse first, and omit the columns last.

This order will also produce the same matrix which is given as an estimate of the mixing matrix a by the FastICA Matlab function [9] [3].

Many authors interested in the inverses of demixing matrices have usually considered the case $m_{\max} = d_{\max}$ only, so this issue has been left without emphasis. [10] [11] [12]

4.6 Summary

We have now proposed four different methods to exploit the posterior (7). The sparse prior with gradient ascent, the sparse prior with an infinitesimal coefficient, the method of local translations of point sources, and the gradient of the likelihood at the origin. We pointed out that the maximization of the posterior is closely related to the objective $f\hat{s} \approx w^+$, which has been studied by many authors in the past, and that the gradient of the likelihood at the origin is closely related to the method published by Hild and Nagarajan. We hope we have now introduced some clarity and unification to this topic by explaining the relationships of the different methods.

5 Simulations

We defined an artificial forward matrix f as explained in the caption of Figure 1. We set $n_{\max} = 100$, $k_{\max} = 1000$, $t_{\max} = 10000$ and $h = 1$. We tested all four methods, which we explained in the theory section, with simple simulations. The order of the four methods here is not the same as in Section 4, since it is convenient to begin with the simplest implementation, which is the gradient of the likelihood at the origin.

In the simulations the statistically independent data c was sampled so that each component obeys a distribution of the form $p(s_m(t)) \sim e^{-|s_m(t)|^\gamma}$ with some coefficient $\gamma < 2$. We used choices $\gamma = 1$ and $\gamma = 1.94$. The reason for the choice close to 2 is that we are interested in investigating the way in which the noise ε disturbs the ICA and the subsequent inverse solution. Thus in some simulations we deliberately want to choose the c in such way that it easily gets lost among the Gaussian noise.

In ICA we always used a preliminary PCA, retaining enough principal components to explain at least 90% of the total variance. The demixing was performed by the FastICA algorithm [9].

Everytime a demixing matrix w was estimated, its rows were normalized so that $\|w_{m*}\| = 1$ for all m. Also the forward matrix f was multiplied with a real coefficient so that the mean of the norms $\|f_{*k}\|$ was one. These details are not important for the discussion, but we mention them since otherwise

[3] http://research.ics.aalto.fi/ica/fastica/code/FastICA_2.5.zip

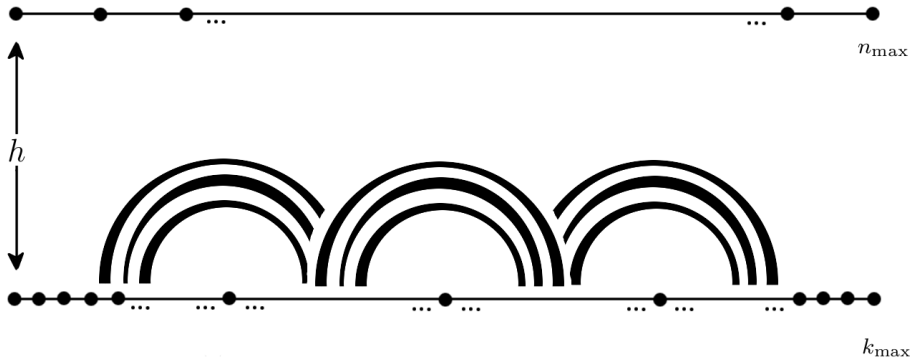

Figure 1: An artistic view of the setting used to define our forward matrix f in simulations. The spatial source space is a one dimensional line divided into $k_{\max}$ points, and $n_{\max}$ measurement sensors are located above it. In the figure the number of localized sources is $m_{\max} = 3$, each source generating a field whose magnitude is proportional to the inverse of the distance. The precise formula for f is $f_{nk} = 1/\sqrt{h^2 + (x_n - x_k)^2}$, where $x_n = -1 + \frac{2(n-1)}{n_{\max}-1}$ and $x_k = -1 + \frac{2(k-1)}{k_{\max}-1}$. This means that the source space can be interpreted as the discretized interval $[-1, 1] \times \{0\}$, and the measurement sensors can be considered to be contained in the set $[-1, 1] \times \{h\}$, with the height h.

the results could not be reproduced with the parameter values mentioned below.

5.1 The gradient of the likelihood at the origin

As the first experiment, we compared the Hild & Nagarajan function to the gradient of the likelihood at the origin. Here we studied only one component, meaning that we had $m_{\max} = 1$. We parametrized the noise covariance Σ with real parameters σ_0 and σ_1. The covariance is defined as a diagonal 100×100 matrix, and its diagonal values are defined by

$$\sqrt{\Sigma_{nn}} = \sigma_0 + (\sigma_1 - \sigma_0)\frac{n-1}{99}. \tag{43}$$

So if $\sigma_0 < \sigma_1$, we will have less noise at the measurement channels on the left (small n), and more noise at the measurement channels on the right (large n). We fixed the true spatial source as $s_{k1} = \delta_{k,250}$. In our experiment, we kept σ_0 at a fixed value 10^{-3}, while σ_1 ran through values $10^{-3}, \ldots, 10^{+1}$ in a loop. With each fixed σ_1, we generated a 1×10000 sample matrix c with $\gamma = 1.94$ as explained in the beginning of Section 5, a 100×10000 Gaussian

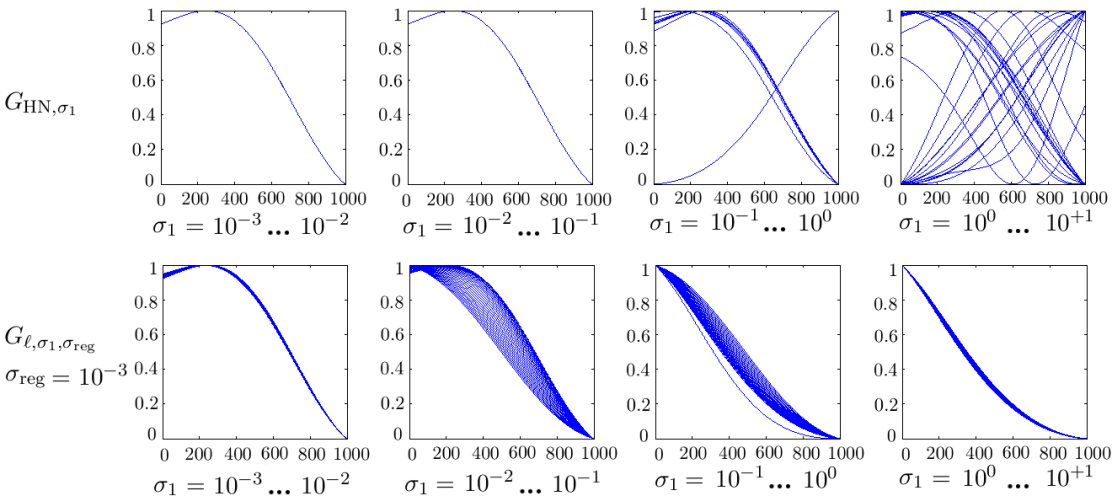

Figure 2: On top row are graphs generated by function G_{HN,σ_1}, and on bottom by $G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}$ with the parameter $\sigma_{\mathrm{reg}} = 10^{-3}$. The functions are defined in Equation (44). Small values of σ_1 are on the left, and large values on the right. Each picture contains 25 graphs with different fixed values of σ_1. In the pictures, the vertical axes denote the values of the functions G_{HN,σ_1} and G_{ℓ,σ_1}, and the horizontal axes denote the spatial space $\{1, 2, \ldots, k_{\max}\}$. With sufficiently small σ_1 (small noise), the maxima are at the correct location $k \approx 250$. With larger σ_1 (greater noise) both functions fail to produce the maxima at $k \approx 250$, but the failures manifest in different ways.

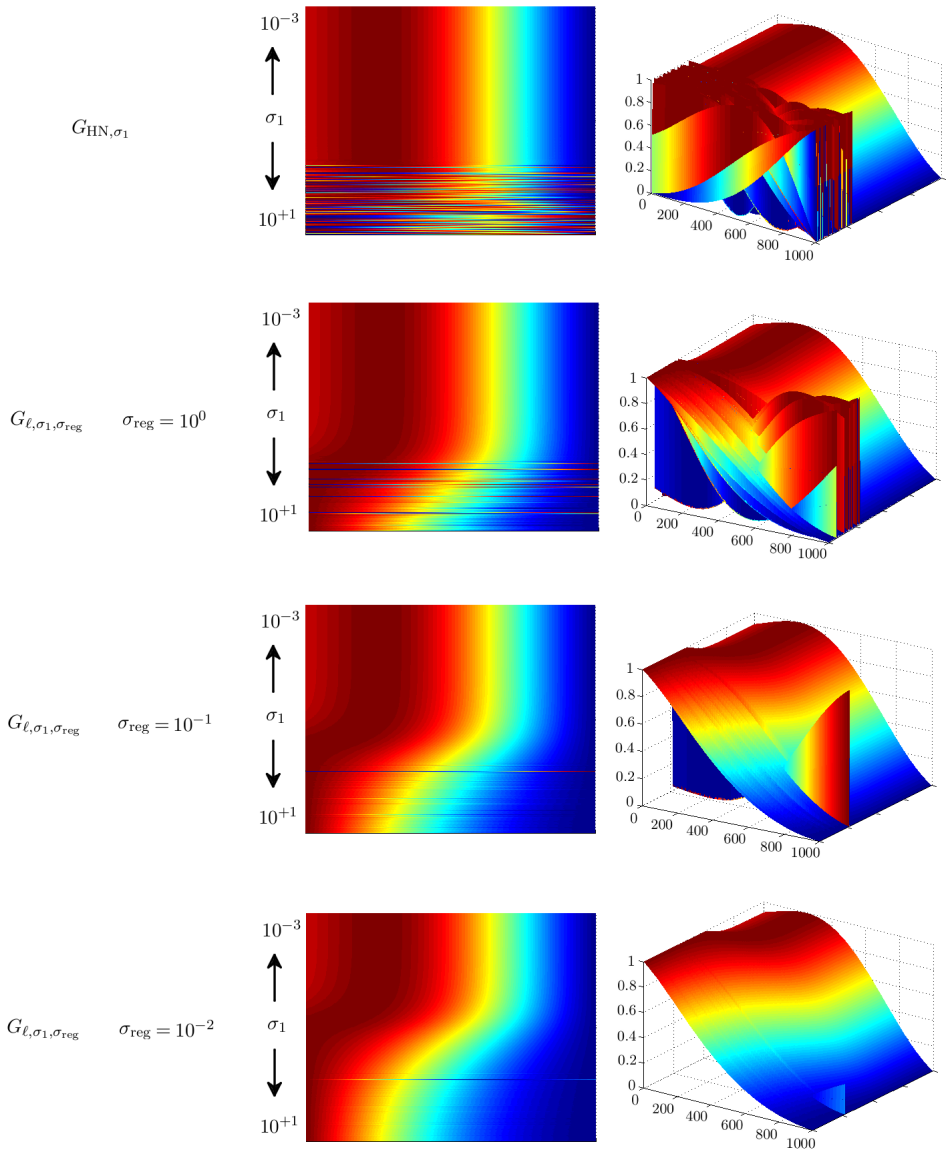

Figure 3: On the top-most row are values of the function G_{HN,σ_1}. The values are the same as those shown on the upper row of Figure 2, but now visualized as a surface parametrized by the parameters k and σ_1. On the rows below are values of $G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}$ with parameters $\sigma_{\mathrm{reg}} = 10^{0}, 10^{-1}, 10^{-2}$. Smaller σ_{reg} tend to produce stronger bias, while larger σ_{reg} leave the results more volatile to the noise.

noise ε, and computed $x = fsc + \varepsilon$. The Gaussian noise was generated so that each column $\varepsilon(t)$ obeys the parametrized covariance Σ. Then a 1×100 demixing matrix w was estimated by the FastICA algorithm. With each fixed w we computed the functions

$$\begin{aligned} G_{\mathrm{HN},\sigma_1}(k) &= \alpha + \beta|\big((w^+)^T\overline{f}\big)_{1k}|, \\ G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}(k) &= \alpha + \beta|\big(wxx^T(\Sigma + \sigma_{\mathrm{reg}}\mathrm{id})^{-1}\overline{f}\big)_{1k}|. \end{aligned} \tag{44}$$

Here the coefficients α and β were chosen so that conditions

$$\begin{aligned} \min(G_{\mathrm{HN},\sigma_1}) = 0, \quad \max(G_{\mathrm{HN},\sigma_1}) = 1, \\ \min(G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}) = 0, \quad \max(G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}) = 1 \end{aligned} \tag{45}$$

held. This means that each function had their own α and β. This way the functions can be plotted simultaneuously and compared nicely. The matrix $\overline{f}$ was defined by a formula $\overline{f}_{*k} = \frac{f_{*k}}{\|f_{*k}\|_2}$. Both of the functions G_{HN,σ_1} and $G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}$ are supposed to tell the location of the point source, which we now know to be the location $k = 250$, so we should hope the approximations

$$\hat{k}_{\mathrm{HN}} := \arg\max_k G_{\mathrm{HN},\sigma_1}(k) \approx 250, \qquad \hat{k}_\ell := \arg\max_k G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}(k) \approx 250 \tag{46}$$

to hold. The function G_{HN,σ_1} is based on the idea by Hild and Nagarajan [13], and the function $G_{\ell,\sigma_1,\sigma_{\mathrm{reg}}}$ is based on the gradient of the likelihood at the origin, $\nabla\ell(s = 0)$, which we explained in Section 4.4. The Σ^{-1} from Equation (39) has now been replaced with a regularized inverse $(\Sigma + \sigma_{\mathrm{reg}}\mathrm{id})^{-1}$, where $\sigma_{\mathrm{reg}} > 0$ is a positive constant, into which we substituted values $\sigma_{\mathrm{reg}} = 10^{-3}, 10^{-2}, 10^{-1}, 1$. The results are shown in Figures 2 and 3, and are discussed in Section 6.

5.2 Sparse prior with gradient ascent

As a second experiment, we investigated the gradient ascent method. We used the same f as in the previous experiment, and this time set $m_{\max} = 3$. The true source s was defined as a 1000×3 matrix by formulas $s_{k1} = \delta_{k,250}$, $s_{k2} = \delta_{k,500}$ and $s_{k3} = \delta_{k,750}$. So each column s_{*m} is a Kronecker delta in its own position.

We defined c as a 3×10000 normalized Laplacian sample ($\gamma = 1$), and ε as a 100×10000 Gaussian sample with a very small covariance. The data $x = fsc + \varepsilon$ was computed, and a demixing matrix w was estimated by the FastICA algorithm.

We then carried out gradient ascent maximization of the objective function $\mathcal{L}_{\mathrm{old}}(s|w^+) - \lambda\|s\|_{1,\varepsilon_{\mathrm{reg}}}$ by using parameters $\lambda = 10^{-1}$ and $\varepsilon_{\mathrm{reg}} = 10^{-3}$. We verified that a gradient step size coefficient $\mu = 0.0021$ caused divergence while $\mu = 0.0020$ did not, and decided that $\mu = 0.001$ is a reasonable choice.

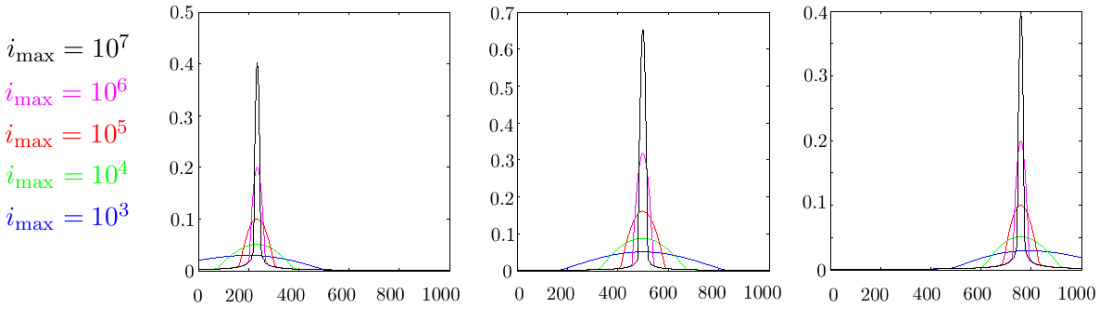

Figure 4: Estimates $s(i_{\max})$ obtained by a gradient ascent. The simulation setting is explained in Section 5.2 text. The plotted components are $s_{*1}(i_{\max})$, $s_{*2}(i_{\max})$ and $s_{*3}(i_{\max})$ from left to right. The colors correspond to different iteration numbers $i_{\max}$. The estimates appear to converge towards correct delta functions, which are $s_{k1} = \delta_{k,250}$, $s_{k2} = \delta_{k,500}$ and $s_{k3} = \delta_{k,750}$.

Also, we verified that other values on the interval $0.001, \ldots, 0.002$ did not produce significantly faster convergence than the choice $\mu = 0.001$. The gradient ascent began from the origin $s(0) = 0$, and estimates obtained by different iteration numbers $i_{\max}$ are shown in Figure 4. We see that the estimates converge towards the correct delta functions, although the amount of required iterations seems to be large. According to further simulations not shown here, in this setting maximizing ℓ instead of $\mathcal{L}_{\text{old}}$ produced practically the same results.

Next, we compared the behaviour of $\mathcal{L}_{\text{old}}$ and ℓ under noise, and used the coefficient $\gamma = 1.94$ while generating c. We parametrized Σ again with parameters σ_0 and σ_1 as shown in Equation (43). Parameter $\sigma_0 = 10^{-3}$ was kept as a constant, and σ_1 ran through values $10^{-3}, \ldots, 10^{+1}$. At each fixed σ_1 we generated x, estimated w, and then solved estimates by gradient ascent by using an iteration amount $i_{\max} = 10^4$. We used the same prior as before ($\lambda = 10^{-1}$ and $\varepsilon_{\text{reg}} = 10^{-3}$), and maximized both $\mathcal{L}_{\text{old}}$ and ℓ. With ℓ, we used a regularized input parameter $\Sigma + \sigma_{\text{reg}}\text{id}$ with a parameter $\sigma_{\text{reg}} = 10^{-1}$. Results are shown in Figure 5, and are discussed in Section 6.

5.3 Infinitesimal prior coefficient

By using very small noise we generated the same x that was used to obtain the results shown in Figure 4, and now tested the method of infinitesimal prior coefficient $\lambda > 0$. We used the same $\varepsilon_{\text{reg}} = 10^{-3}$ in the prior $\|s\|_{1,\varepsilon_{\text{reg}}}$ as before. In the analytic formula for $\hat{s}^0$ and the projection matrix we used the regularizing parameter n_{reg} whose definition was explained in Section 4.2. It turned out that surprisingly small n_{reg} work very well.

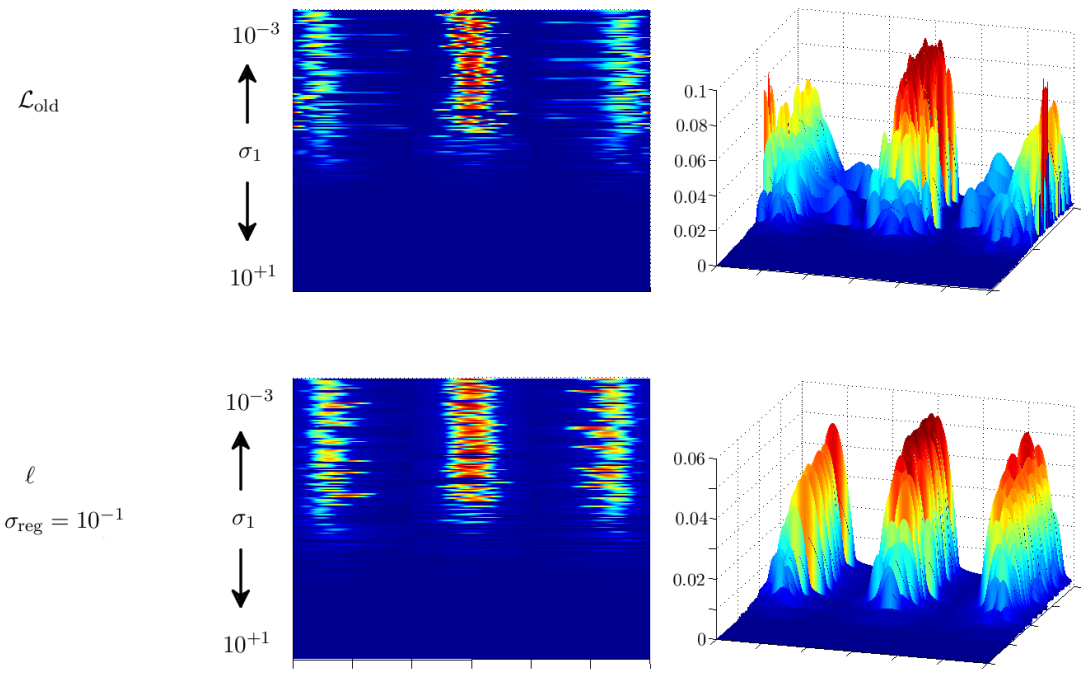

Figure 5: Spatial estimates obtained by gradient ascent under noise. The simulation setting is explained in Section 5.2 text. All three components are plotted on the same axis consisting of 3000 points. The vertical axis (in two dimensional visualization on left) denotes the noise parameter σ_1. With sufficiently small noise we obtain bumps in roughly the correct spatial locations. The bumbs obtained by maximizing $\mathcal{L}_{\text{old}}$ appear to be more noisy.

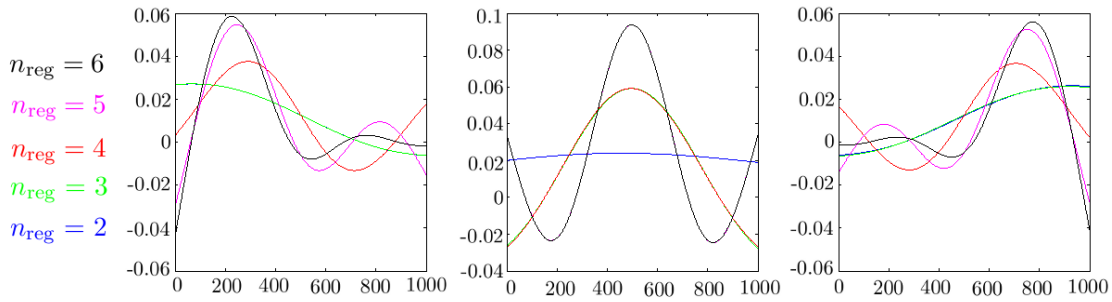

Figure 6: Spatial estimates obtained by the formula $\hat{s} = f^{+}_{n_{\text{reg}}} w^{+}$. Different colors correspond to different values for the regularization parameter n_{reg}, whose definition was explained in Section 4.2.

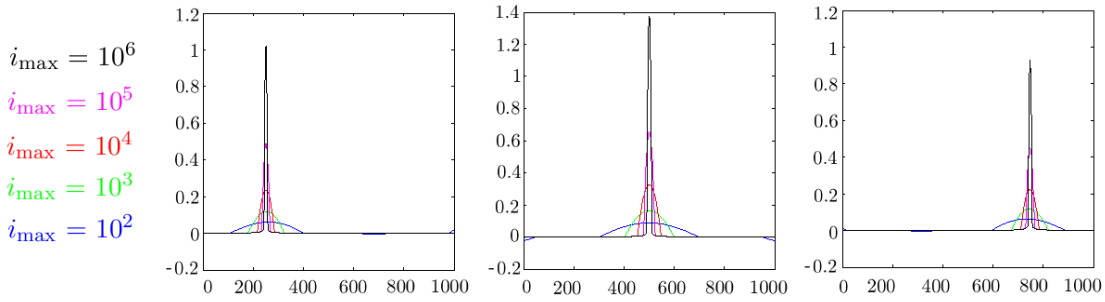

Figure 7: Spatial estimates obtained by maximizing $\mathcal{L}_{\text{old}}(s|w^+) - \lambda\|s\|_{1,\varepsilon_{\text{reg}}}$ with an infinitesimal $\lambda > 0$. This means that we have maximized the function $-\|s\|_{1,\varepsilon_{\text{reg}}}$ with a gradient ascent and with a constraint as explained in Section 4.2. The value $n_{\text{reg}} = 4$ was used in the initial point $s(0) = f^+_{n_{\text{reg}}} w^+$, and estimates $s(i_{\max})$ are shown with different iteration numbers $i_{\max}$.

Some estimates obtained by a formula $\hat{s}^0 = f^+_{n_{\text{reg}}} w^+$ are shown in Figure 6. Intuitively these estimates appear to be linear combinations of some low frequency Fourier components. These estimates can have several local maxima, but the global maxima are reasonably close to the true source locations. Sparse estimates obtained by a gradient ascent are shown in Figure 7, and are discussed in Section 6.

In the gradient ascent we used the step size coefficient $\mu = 0.001$. With this method choosing a proper step size is more difficult than with the ordinary gradient ascent, because it seems that surprisingly large values can be substituted into μ without obvious divergence to infinity. Instead of such divergence, too large μ usually results in bad oscillations and extremely slow convergence or divergence. In this setting values $\mu = 0.1, \ldots, 0.05$ produced messy estimates with significant high frequence components. The value $\mu = 0.01$ appeared to work, and we considered the value $\mu = 0.001$ reasonable and safe.

5.4 The method of translating point sources

As the last experiment we carried out simulations with the method of translating point sources, which was explained in Section 4.3. Here $m_{\max}$ ran through values $1, 2, 3$. Again we parametrized the covariance Σ with parameters σ_0 and σ_1, and with a formula $\sqrt{\Sigma_{nn}} = \sigma_0 + (\sigma_1 - \sigma_0)\frac{n-1}{99}$. We kept $\sigma_0 = 10^{-3}$ at a fixed value, and allowed σ_1 to run through 1000 values on the interval $10^{-3}, \ldots, 10^{+1}$. With each fixed σ_1, source locations $k_1, \ldots, k_{m_{\max}}$ were sampled from a distribution with weight on the left by roughly the relation $p(k) \sim \frac{1}{\sqrt{k}}$. This was achieved by sampling real numbers $\tilde{k}_1, \ldots, \tilde{k}_{m_{\max}}$

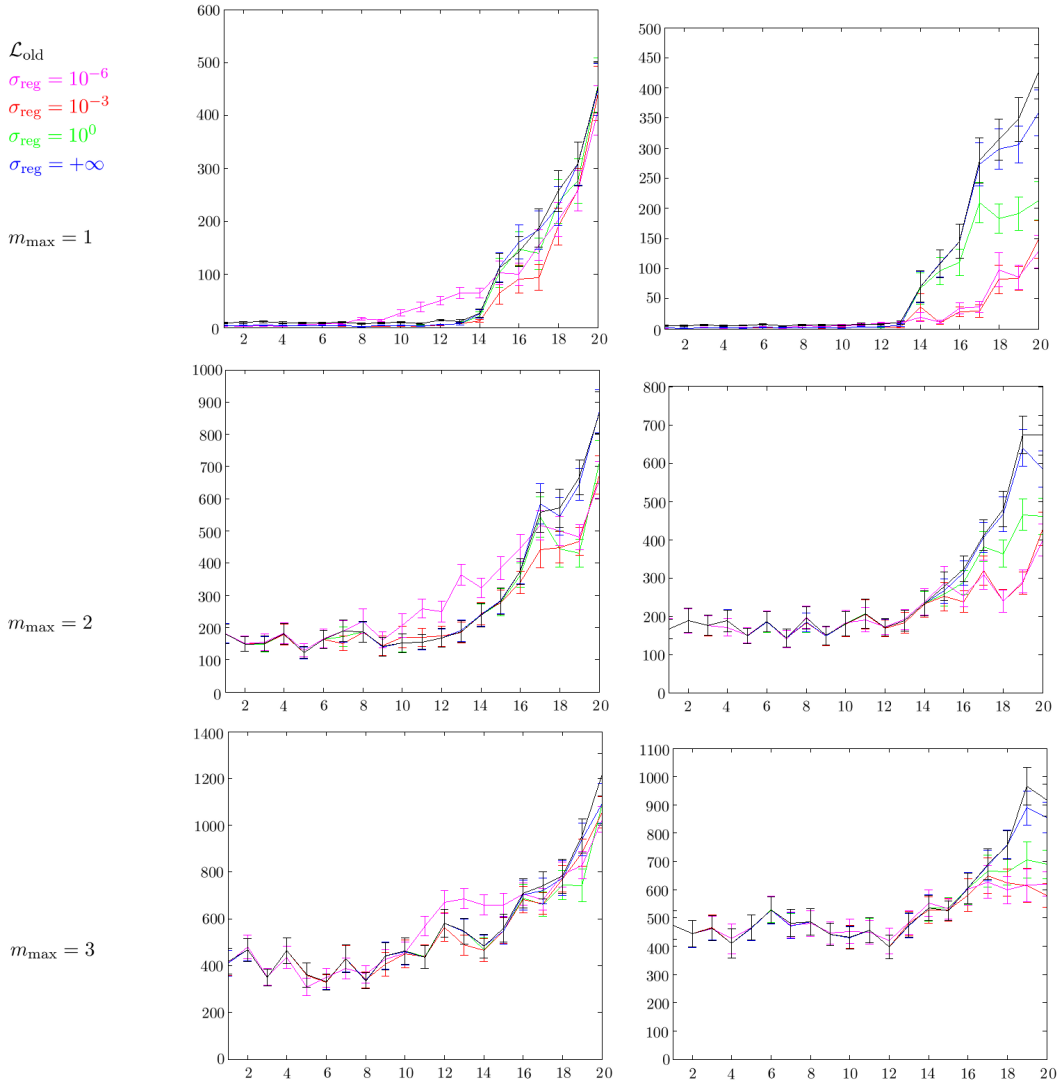

Figure 8: The horizontal axes describe the parameter σ_1 which goes through the values $10^{-3}, \ldots, 10^{+1}$ (resolution of 1000 points). The vertical axes denote the values of the errors $E_{\mathcal{K}}$ defined in (47). On left the noise was generated to be significant with large n, and on right so that the covariance matrix itself was random, as explained in the Section 5.4 text. The horizontal axes have been divided into 20 blocks, and in each a mean of the corresponding 50 values has been plotted. The widths of the bars have been computed by a formula $\frac{2}{\sqrt{50}}\text{Var}$. Black graphs show the results obtained by maximizing $\mathcal{L}_{\text{old}}$, and the colored graphs show the results obtained by maximizing ℓ. Different colors correspond to different parameters σ_{reg}.

from the uniform distribution on $[0,1]$, and setting $k_m = [999 \cdot \tilde{k}_m^2] + 1$ (here $[\cdot]$ denotes the floor function). The amplitudes $a_1, \ldots, a_{m_{\max}}$ were sampled from a uniform distribution on the interval $[\frac{1}{2}, \frac{3}{2}]$, and the spatial source was then computed by a formula $s_{km} = a_m \delta_{kk_m}$. The data c, with a parameter $\gamma = 1.94$ as explained in the beginning of Section 5, and the Gaussian noise ε were generated, and the measurement data $x = fsc + \varepsilon$ was computed.

Then the demixing matrix w was estimated by the FastICA algorithm, and estimates $\hat{k}_1, \ldots, \hat{k}_{m_{\max}}$, $\hat{a}_1, \ldots, \hat{a}_{m_{\max}}$ were solved in various ways. One set of estimates was computed by maximizing $\mathcal{L}_{\text{old}}(s|w^+)$, where s was parametrized by $s_{km} = \hat{a}_m \delta_{k\hat{k}_m}$. Other sets of estimates were computed by maximizing $\ell(s|w, x, \Sigma + \sigma_{\text{reg}}\text{id})$, where σ_{reg} is a real parameter, which was given various values. Finally quantities

$$E_{\mathcal{K}} = \sum_{m=1}^{m_{\max}} |\hat{k}_m - k_m|, \tag{47}$$

were computed to measure the success of the inverse solution. The components were permutated to minimize $E_{\mathcal{K}}$ as a last step. We expect $E_{\mathcal{K}}$ to be close to zero, if the inverse solutions are successful. We used 10 different random initial values for k_m and a_m, and chose those which resulted in the largest ℓ or $\mathcal{L}_{\text{old}}$ after maximization. The prior log $p(S = s)$, which forces s to consist of point sources, is not concave, and a large amount of random initial points is one natural solution to the local maxima problem. The results are shown in Figure 8 on left.

We then repeated the point source translation experiment with a different noise. Again, $\sigma_0 = 10^{-3}$ was fixed, and σ_1 ran through 1000 values on the interval $10^{-3}, \ldots, 10^{+1}$, but the covariance was defined as follows. With each fixed σ_1, we sampled a set of coefficients $\beta_1, \ldots, \beta_{n_{\max}}$, each independently from the uniform distribution on the interval $[0,1]$. Then the covariance was defined by a formula $\sqrt{\Sigma_{nn}} = \sigma_0 + (\sigma_1 - \sigma_0)\beta_n$. Again the noise increases with the growing parameter σ_1, but the increase occurs in a more random manner, not weighted for large n particularly. The true locations of the sources were now sampled from the uniform distribution on the spatial space $\{1, 2, \ldots, 1000\}$. This time the nature of the bias, produced by the Σ^{-1} term in the likelihood, is not intuitively clear. The results are shown in Figure 8 on right.

6 Discussion

Our study has been focused on studying the relationship of the functions $\mathcal{L}_{\text{old}}$ and ℓ. In Sections 2.2 and 3 we explained theoretically that if the noise covariance is proportional to the identity, the maxima of $\mathcal{L}_{\text{old}}$ and ℓ are the same. The question about how the maxima are related when the noise

covariance is not proportional to the identity was left open in the theoretical analysis. Ultimately, we have been unable to give a simple answer to this question, since the relationship of the maxima of $\mathcal{L}_{\text{old}}$ and ℓ seems to be complicated in general, and seems to be different with different kinds of noise. In Section 5 we approached this question in an obvious manner by constructing artificial situations were the noise covariance was intentionally not proportional to the identity, and investigated empirically what happens. Next we discuss these results and the four methods that we used in more detail.

Translating point sources as an evaluation platform The method of translating point sources is problematic for several reasons, and we do not recommend it very strongly. One problem is that it is not very elegant and its implementation leads to more programming work than the implementation of the other methods. It can suffer from the problem of local maxima, which is not fatal, but can be a nuisance. If the real source distribution contains several spikes for each individual component m, the method we used should be generalized to support the several point sources, but this would lead to further problems which we leave for future research.

However, this method has one advantage over the other methods, and it is that if we know in advance the correct results, it is straigthforward to estimate the quality of the inverse solution by simply computing the distance $|\hat{k} - k|$ of the correct spike location and its estimate. For this reason, the method of translating point sources turned out to be very practical in our study. With other methods the results cannot be measured with equal precision. For example, one obvious way to measure the quality of some distribution estimate $\hat{s}$ would be to use some vector (or matrix) norm $\|\hat{s}-s\|$. Most obvious norms would be very bad since they do not take into account the physical distances of the point sources in the space $\{1, 2, \ldots, k_{\max}\}$.

Although we were unable to give a simple answer to the relationship of $\mathcal{L}_{\text{old}}$ and ℓ in general, we were able to prove, by using the method of translating point sources, that in some circumstances maximizing ℓ produces more accurate results than maximizing $\mathcal{L}_{\text{old}}$. This was proved in Figure 8 where we see that the colored graphs are at some points clearly below the black graphs, meaning that maximizing ℓ has produced statistically smaller errors than maximizing $\mathcal{L}_{\text{old}}$. However, we also see that sometimes the violet graph is above the black graph. Apparently using too small regularizing parameter σ_{reg} can result in inferior results.

We were unable to obtain equally clear results with the three other methods, but at this point we propose the following reasoning. We have proven, by using the method of translating point sources, that sometimes maximizing ℓ produces better results than maximizing $\mathcal{L}_{\text{old}}$. Hence, we have some reason to believe that maximizing ℓ could produce better results than max-

imizing $\mathcal{L}_{\text{old}}$ with the other methods too.

Connection between $\mathcal{L}_{\text{old}}$ and ℓ The gradient of the likelihood at the origin and the related Hild-Nagarajan method have the advantage that the methods are very fast, since they do not use any iterative processes. The other side of the coin is that the results by these methods are crude and do not contain the same amount of information as the outputs of the other methods. For these reasons these two methods could be considered as the quick and crude checks that preceed the application of more advanced methods. This way we get some idea what to expect of the inverse solutions.

In Section 4.4 we gave a theoretical proof for the result that the Hild-Nagarajan function is equivalent with the gradient $\nabla\ell(s = 0)$ at the origin, when the noise is small. Figures 2 and 3 verify this result, since with small σ_1 the shapes of the graphs by both G_{HN,σ_1} and $G_{\ell,\sigma_1,\sigma_{\text{reg}}}$ are very similar. It seems that there exists some critical value for σ_1, and once σ_1 exceeds it, the w starts to exhibit random behaviour, and consequently G_{HN,σ_1} starts to produce rather random results too. The method proposed by us attempts to deal with the noise in a more logical manner. It seems to sense that since there is significant noise in the direction of large n, consequently the w might be drawn towards this direction of greater noise, and hence the estimate $\hat{k}$ might be drawn towards right end of the spatial space too (large k). Then $G_{\ell,\sigma_1,\sigma_{\text{reg}}}$ introduces a logical counter bias, which pushes the estimate $\hat{k}$ to left (small k). It must be admitted though that this counter bias is not sufficiently precise to maintain the correct result $\hat{k} \approx 250$.

In fact, in our studies, we never encountered a situation where the Hild-Nagarajan method or the maximization of $\mathcal{L}_{\text{old}}$ would catastrophically fail, while the maximization of the likelihood ℓ or the examination of its gradient at the origin would nicely succeed. So it seems that the difference between $\mathcal{L}_{\text{old}}$ and ℓ is not so dramatic. Considering the data shown in Figure 8, it seems that under difficult noise using ℓ can improve the probability that the estimate has the right shape, and can reduce the expected error in the location of the distribution maximum. Thus the relationship of $\mathcal{L}_{\text{old}}$ and ℓ should be seen to be of this nature only.

Using sparse priors Comparing Figures 4 and 7 we see that the gradient ascent with the infinitesimal λ can produce estimates that are very similar to those produced by the ordinary gradient ascent. A close look at the figures reveals that the infinitesimal λ produced sharper estimates with a smaller amount of iterations. Also the infinitesimal λ brings the advantage that we do not need to go through a trouble of finding a suitable value for a real parameter λ. In light of these remarks the method of infinitesimal λ appears to be very interesting.

The method of infinitesimal λ introduces a new parameter n_{reg} which

was not present in the ordinary gradient ascent. This parameter is easier to deal with than the real λ, since n_{reg} takes integer values, and small values usually work.

In ordinary gradient ascent we used a parameter value $\lambda = 10^{-1}$. This choice was made because we found empirically that with larger values such as $\lambda = 1, 10, \ldots$ the prior term would start to dominate the $\mathcal{L}_{\text{old}}$ term excessively. Here we could check this easily because we knew what the estimates were supposed to look like. On the other hand, using smaller values such as $\lambda = 10^{-2}, 10^{-3}, \ldots$ would have resulted in slower convergence.

In Figure 5 we see how the gradient ascent works with both $\mathcal{L}_{\text{old}}$ and ℓ under noise. It seems that there exists some critical value for σ_1, and the methods work while σ_1 is below it. There is no visible difference in the results obtained by $\mathcal{L}_{\text{old}}$ and ℓ with respect to the question of the critical σ_1, although it is visible in the figure that the estimates obtained by $\mathcal{L}_{\text{old}}$ are slightly more random.

On regularization of the covariance matrix A significant source of trouble, with methods that involve adjustable input parameters, is that there may be no obvious way to decide the optimal values for them. Once a guess has been made for the parameters, the results can turn out nonsensical. With real inverse problems, where the correct answers are not known in advance, this issue can become a frustrating concern. To end our discussion here we propose one possible strategy to deal with the regularization of the covariance matrices. In simulations we mentioned the regularized covariance parameter $\Sigma + \sigma_{\text{reg}}\text{id}$, where $\sigma_{\text{reg}} > 0$. In fact also the matrix $\frac{1}{t_{\max}} xx^T$ can be regularized by replacing it with $\frac{1}{t_{\max}} xx^T + \sigma'_{\text{reg}}\text{id}$, although we did not use this in our simulations. We have theoretical reasons to believe that with very large parameters σ_{reg} and σ'_{reg} the regularized ℓ should have the same maxima as $\mathcal{L}_{\text{old}}$. This remark implies the following three step strategy. As a first step, obtain inverse solutions by using $\mathcal{L}_{\text{old}}$. So in the first step, we don't use the information in the data covariance or the assumed noise covariance. As a second step, verify that we can recover the same inverse solutions with the likelihood ℓ by using very heavy regularization in the covariance matrices (large values for σ_{reg} and σ'_{reg}). Once the second step has been checked, the final third step is to observe what happens to the inverse solutions when the values of the regularization parameters are brought downwards. If the inverse solutions go through some reasonable changes, it could be that they are being enhanced by the information from the covariance matrices. This way the covariance information can be exploited with a reasonable safety.

7 Conclusions

The idea of using ICA as a tool to solve spatial inverse solutions has received some attention in the past [10] [11] [12] [13], but the topic has not yet been approached in a systematic and principled manner. In this paper we showed how to solve a Bayesian posterior for the spatial source distribution, by using an ICA estimate as an input. This enabled us to derive new inverse solution algorithms and to better understand algorithms already published by others. We demonstrated in a controlled artificial setting, that in some circumstances our new methods produce statistically more accurate inverse solutions than the existing methods.

The mathematical theory we have presented here is not directly related to any particular physical application, but historically this direction has attracted attention in the brain imaging community in particular. The same approach could also be used in other inverse problems where ICA is possible.

References

[1] M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, O. V. Lounasmaa: Magnetoencephalography—theory, instrumentation and applications to noninvasive studies of the working human brain. Reviews of Modern Physics vol. 65, num. 2, pages 413-497, 1993

[2] S. Baillet, J. C. Mosher, R. M. Leahy: Electromagnetic Brain Mapping. IEEE Signal Processing Magazine, 2001

[3] A. Hyvärinen, J. Karhunen, E. Oja: Independent Component Analysis. Wiley-Interscience, 2001

[4] P. Comon: Independent component analysis—a new concept? Signal Processing, vol. 36, pages 287–314, 1994

[5] S. Makeig and T.-P. Jung and A. J. Bell, D. Ghahramani and T. Sejnowski: Blind separation of auditory event-related brain responses into independent components. Proc. National Academy of Sciences (USA), vol. 94, pages 10979–10984, 1997

[6] R. Vigário, J. Särelä, V. Jousmäki and M. Hämäläinen and E. Oja: Independent Component Approach to the Analysis of EEG and MEG Recordings. IEEE Trans. Biomedical Engineering, vol. 47, num. 5, pages 589-593, 2000

[7] Brookes, M.J. and Woolrich, M.W. and Luckhoo, H. and Price, D. and Hale, J.R. and Stephenson, M.C. and Barnes, G.R. and Smith, S.M. and Morris, P.G.: Investigating the electrophysiological basis of resting state networks using magnetoencephalography. Proc. National Academy of Sciences (USA), vol 108, pages 16783–16788, 2011

[8] Ramkumar, P. and Parkkonen, L. and Hari, R. and Hyvärinen, A.: Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. Human Brain Mapping, vol. 33, num. 7, pages 1648-1662, 2012

[9] A. Hyvärinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10(3):626-634, 1999

[10] L. Zhukov, D. Weinstein, C. Johnson: Independent Component Analysis for EEG Source Localization. IEEE Engineering in Medicine and Biology Magazine, vol. 19, iss. 3, pages 87-96, 2000

[11] Y. Chen, Q. Zhang, Y. Kinouchi: Blind Source Separation by ICA for EEG Multiple Sources Localization. IFMBE Proceedings, 2007, Volume 14, Part 16, 2760-2763. 2007

[12] A.C. Tang, B.A. Pearlmutter, N.A. Malaszenko, D.B. Phung, B.C Reeb: Independent Components of Magnetoencephalography: Localization. Neural Computation 14, 1827-1858, 2002

[13] K.E. Hild, S.S. Nagarajan: Source Localization of EEG/MEG Data by Correlating Columns of ICA and Lead Field Matrices. IEEE Trans Biomed Eng. 2009 Nov; 56(11):2619-26. 2009

[14] K. H. Knuth: "Bayesian Source Separation and Localization" in Bayesian Inference for Inverse Problems, San Diego, CA, Proceedings of the SPIE Series. Bellingham, WA: SPIE, vol. 3459, Jul. 1998, pp. 147-158, 1998

[15] S. Roberts, R. Choudrey: "Bayesian Independent Component Analysis with Prior Constraints: An Application in Biosignal Analysis" in Deterministic and Statistical Methods in Machine Learning, vol. 3635, Berlin, Germany. Springer-Verlag 2005. pp. 159-179, 2005

[16] K. Zhang and H. Peng and L. Chan and A. Hyvärinen: ICA with Sparse Connections: Revisited. Proc. Int. Conference on Independent Component Analysis and Blind Signal Separation (ICA2009), pages 195–202, 2009

[17] Stone, J. V. and Porrill, J. and Porter, N. R. and Wilkinson, I. D.: Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. NeuroImage, vol. 15, num. 2, pages 407–421, 2002

[18] K. Uutela, M. Hämäläinen, and E. Somersalo. Visualization of magnetoencephalographic data using minimum current estimates. NeuroImage, 10:173-180, 1999.

[19] Mosher, J. C. and Lewis, P. S. and Leahy, R. M.: Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Transactions on Biomedical Engineering, vol. 39, num. 6, pages 541–557, 1992.