REGULARIZED SECOND ORDER SOURCE SEPARATION

I. Schießl[1]*,H. Schöner*[1]*, M. Stetter*[1]*, A. Dima*[1] *and K. Obermayer*[1]

[1] Dept. of Computer Science, Technical University of Berlin,
D-10587 Berlin, Germany; email: ingos@cs.tu-berlin.de

ABSTRACT

The task of separating signals from experimentally measured linear mixtures is often complicated by the presence of noise sensor noise and statistical dependencies between the original sources, which often makes standard independent component analysis (ICA) algorithms fail [1, 2]. One way to overcome these problems is to introduce additional knowledge we have about the mixing process and the signals themselves.

Here we suggest to add a regularization term to the cost function of multishift extended spatial decorrelation (multishift ESD, [2]) which contains prior information about the time-course of one or more original sources. Using an artificial toy dataset and a dataset that contains prototype signals obtained from optical recording of brain activity we show that the regularization term improves the separation results at different noise levels.

1. INTRODUCTION

The basic assumptions underlying standard independent component analysis (ICA) are, that the original sources are statistically independent and that the mixing process itself is linear. A lot of different separation algorithms have been developed and were proven to be successful as long as the signals fulfill the above assumptions.

Real data, however, satisfy the "independence" and the "linearity" assumptions only approximately. One common problem is "sensor" noise. In contrast to the so called "source" noise, which is treated as an additional source signal in the demixing process, "sensor" noise is added to the measured signals after the mixing process. "Sensor" noise may originate from the read-out noise of a CCD chip in a video camera, from photon shot noise, from noise within individual microphones, etc.

Another problem are small statistical dependencies between sources. If sources are – for example – evoked by the same event, dependencies or convolutive effects may be introduced into the measurements. Our main interest in the application of ICA algorithms is the separation of intrinsic signals that arise from neural activity in optical imaging experiments. In optical imaging, a cortical area of interest is illuminated with monochromatic light of wavelengths usually between 500 - 800 nm. This area is then recorded with a sensitive CCD- or video camera. Changes in reflectance of this light from the cortex are mainly due to variations in the light scattering properties of the tissue and to variations in the local concentrations of deoxygenated and oxygenated hemoglobin. Typically these changes are very small and do not exceed 0.1% of the reflected light[3]. Because of these small intensities the signal to noise ratio in optical imaging experiments is around 1 (0 dB). Statistical dependencies are a problem in this data, because different signals can be evoked by the same stimulus.

In [1, 4] we introduced a method called extended spatial decorrelation (ESD), that was derived from an algorithm suggested by Molgedey and Schuster [5], and which is implicitly based on the assumption that sources are spatially smooth. In this work we show how the performance of ESD on data with properties similar to those of optical imaging recordings can be enhanced by using prior knowledge about the time course of the sources in a regularization framework. For the comparison of the separation quality at different noise levels we used one artificial dataset of three spatially smooth sources (figure 1 top row) and one dataset of three prototype patterns from optical imaging recordings (figure 1 bottom row).

This work was supported by DFG OB102/2-1 and Wellcome Trust 050080/Z/97

2. ALGORITHM

Let m be the number of mixtures (here observed image frames), $\mathbf{r}$ the sample index, i.e. a vector specifying a pixel in the image data set, and R the total number of pixels. In equation (1) the observation vectors $\mathbf{y}(\mathbf{r}) = (y_1(\mathbf{r}), \ldots, y_m(\mathbf{r}))^T$ are assumed to be linear mixtures of n unknown sources $\mathbf{s}(\mathbf{r}) = (s_1(\mathbf{r}), \ldots, s_n(\mathbf{r}))^T$

$$\mathbf{y}(\mathbf{r}) = \mathbf{A}\mathbf{s}(\mathbf{r}) + \eta \qquad (1)$$

with $\mathbf{A}$ being the $m \times n$ mixing matrix and η describing the sensor noise.

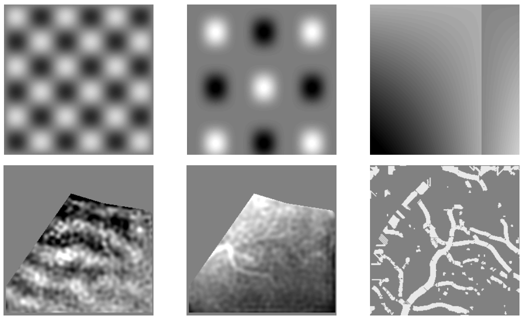


Figure 1: The two datasets used for the following benchmarks. The sources in the top row are referred to as "smooth" sources; The sources in the bottom row are referred to as "natural" sources. For description see section 3

The goal of ICA is to obtain optimal source estimates $\hat{\mathbf{s}}(\mathbf{r})$ under the assumption that the original sources are independent. The ESD approach, on the other hand, uses only second order statistics, by minimizing cross-correlation between (spatially shifted) sources; the underlying assumptions are that sources are uncorrelated and spatially smooth.

In the noiseless case $\mathbf{W} = \mathbf{A}^{-1}$ would be the optimal demixing matrix. In presence of sensor noise (which is added after the mixing), however, $\mathbf{W}$ also has to compensate for the added noise: $\hat{\mathbf{s}}(\mathbf{r}) = \mathbf{W}\mathbf{y}(\mathbf{r}) = \mathbf{W}\mathbf{A}\mathbf{s}(\mathbf{r}) + \mathbf{W}\eta$. So even with perfect prior knowledge the optimal $\mathbf{W}$ can deviate from $\mathbf{A}^{-1}$, to compensate for noise. BSS algorithms are generally only able to recover the original sources up to a permutation and scaling.

2.1. Extended Spatial Decorrelation

Extended Spatial Decorrelation (ESD) exploits the second order statistics of the observations to find the source estimates. If sources are statistically independent then all source cross-correlations

$$C_{ij}^{(s)}(\Delta\mathbf{r}) = \langle s_i(\mathbf{r}) s_j(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}} \qquad (2)$$
$$= \frac{1}{R}\sum_{\mathbf{r}} s_i(\mathbf{r}) s_j(\mathbf{r}+\Delta\mathbf{r}), \qquad \text{where } i \neq j$$

must vanish for all shifts $\Delta\mathbf{r}$, while the autocorrelations ($i = j$) of the sources remain. This approach for BSS is best suited for sources which are known to be spatially smooth, but whose probability distributions may be unknown. Source separation, i.e. the estimation of the demixing matrix $\mathbf{W}$, is performed by minimizing the cost function

$$E_S(\mathbf{W}) = \sum_{\Delta\mathbf{r}} \sum_{i\neq j} \left(\left(\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T\right)_{i,j} \right)^2 \qquad (3)$$
$$= \sum_{\Delta\mathbf{r}} \sum_{i\neq j} \langle \hat{s}_i(\mathbf{r}) \hat{s}_j(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}}^2$$
$$\text{where } C_{ij}(\Delta\mathbf{r}) = \langle y_i(\mathbf{r}) y_j(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}} \qquad (4)$$

are the correlations of the observed mixtures and $\hat{\mathbf{s}}_i$ are the estimated sources.

For the case of only two shifts the cost function $E_S(\mathbf{W})$ can be minimized very efficiently by solving an Eigenvalue problem (see [5, 1, 4]). Though computationally very efficient, solutions are very sensitive to the choice of shifts used.

An alternative is to use several shifts $\Delta\mathbf{r}$, which leads to a reduction of noise and makes solutions robust. Equation (3) has then to be minimized using an iterative procedure, for which a modified conjugate gradient method, using an adaptive step-size instead of a line search, promised to be a useful method (method in [6] and implementation in [2]). The parameters $\mathbf{W}$ were initialized randomly from a Gaussian distribution with mean 0 and variance 1.

It also turned out to be advantageous to sphere the data ($\mathbf{y}'(\mathbf{r}) = \mathbf{D}\mathbf{y}(\mathbf{r})$, $\mathbf{D} = \langle \mathbf{y}(\mathbf{r})\mathbf{y}^T(\mathbf{r})\rangle_{\mathbf{r}}^{-1/2}$) minimizing equation (3). Because sensor noise is assumed to have no autocorrelation for shifts $\neq (0,0)$, it is even more advantageous to use the sphering method proposed in [7], which uses $\mathbf{D} = \langle \mathbf{y}(\mathbf{r})\mathbf{y}^T(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}}^{-1/2}$, with $\Delta\mathbf{r}$ being small. $\Delta\mathbf{r}$ should be chosen such that as much as possible information about the correlation structure is preserved, while cancelling out the (not correlated) noise (see also [2]).

In the benchmarks presented later in this paper we compare the ESD algorithm using two shifts $\Delta\mathbf{r} \in \{(0,0),(5,5)\}$, and the multishift algorithm (using improved sphering) with the method described in the following subsection.

2.2. Regularization

We now add a regularization term to the cost function, which (1) incorporates prior knowledge about the time course of some or all of the sources and (2) breaks the permutation symmetry in the ordering of the estimated sources. Because $\mathbf{W}^{-1}$ should be close to $\mathbf{A}$ after the decorrelation process, we introduce a regularization term, which punishes deviation of $\mathbf{W}^{-1}$ from $\mathbf{A}$. As column j of $\mathbf{A}$ represents the time course of the source j, we weight this deviation by a regularization parameter α_j whose value reflects the confidence we have in our prior knowledge about source j, and we obtain

$$E_{\mathbf{W}}(\mathbf{W}) = \sum_j \alpha_j \cdot \sum_i \left((\mathbf{W}^{-1})_{ij} - A_{ij}\right)^2 \qquad (5)$$

Altogether we get a cost function

$$E(\mathbf{W}) = E_S(\mathbf{W}) + E_{\mathbf{W}}(\mathbf{W}), \tag{6}$$

which is minimized using the gradient descent procedure mentioned in section 2.1. In the following we compare the two ESD variants described in section 2.1 to this algorithm; first prior knowledge about all sources is used, and later the case with only one regularized source is shown.

3. BENCHMARKS FOR ARTIFICIAL DATA

For testing the performance of the regularized ESD we used two different types of toy datasets. The first dataset (figure 1, top row) consists of three spatially smooth artificial images, that are weakly correlated, similar to the datasets obtained in optical recording experiments. The sources have a variance of 1, and the largest cross-correlation is about 0.1. This dataset is referred to as the "smooth" dataset.

In the second toy dataset (figure 1, bottom row) we used three plausible prototype patterns, obtained by performing multishift ESD analysis on datasets from a real optical imaging experiment. The first image shows an ocular dominance pattern from the striate cortex of a macaque and is a prototype pattern for a stimulus specific response, i.e. the signal we want to separate from the others. The second image shows the response of the capillary bed to the stimulation (the so-called global response pattern). The third image displays the blood vessel pattern extracted from the original data. This dataset is referred to as the "natural" one.

In the case of our spatial analysis each column of the mixing matrix **A** contains the time course of the individual source. We designed two different mixing matrices. The first matrix **A1** contains three time points for each of the three sources and therefor is square. The second mixing matrix **A2** contains ten time points for each of the three sources and is a 10×3 matrix. After the mixing process with **A1** we had three and ten mixtures for **A1** and **A2**, respectively. To each of the mixtures random white noise with varying variance was added to obtain signal to noise ratios between about 15 and 0 dB. In order to score the performance of the separation procedures we calculated the average reconstruction error (RE, [8])

$$\mathsf{RE}(\mathbf{W}) = \mathrm{od}\Big(\sum_{\mathbf{r}} \hat{\mathbf{s}}(\mathbf{r})\mathbf{s}^T(\mathbf{r})\Big), \tag{7}$$

$$\mathrm{od}(\mathbf{C}) = \frac{1}{N}\sum_i \frac{1}{N-1}\Big(\sum_j \frac{|C_{i,j}|}{\max_k |C_{i,k}|} - 1\Big) \tag{8}$$

between the estimated and the original sources.

The correlation between the real and the estimated sources (the argument to "od"), should be close to a permutation matrix, if the separation is successful. If the maxima of two rows are in the same column, the separation is labeled unsuccessful. Otherwise, the normalized absolute sum of non-permutation (cross-correlation) elements is computed and returned as the reconstruction error.

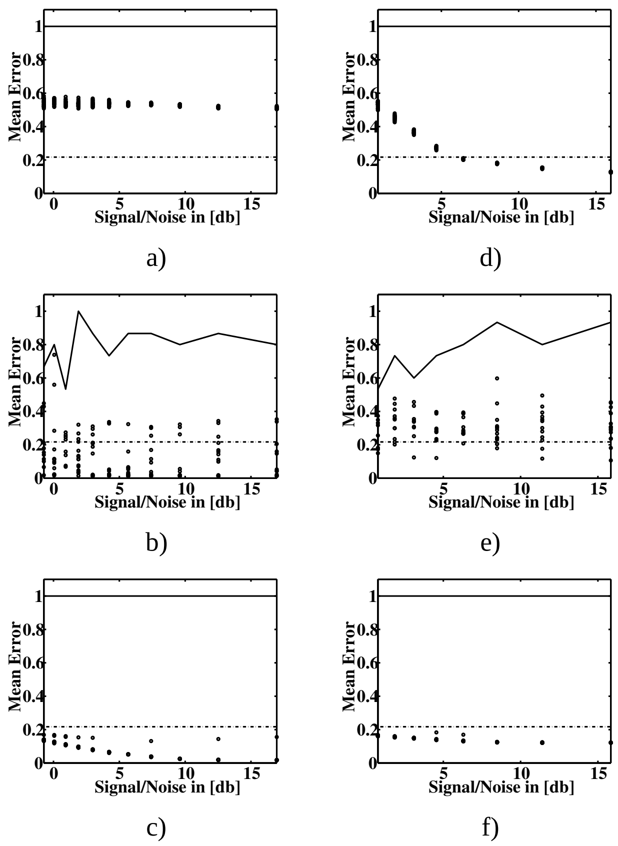


Figure 2: Mean reconstruction error as a function of the signal to noise ratio in dB (15 trials per noise level) for the mixing matrix **A1**. The left column shows the results of the separation for the smooth and the right column for the natural sources. The first row shows the results for ESD with the two shifts $\Delta\mathbf{r} \in \{(0,0),(5,5)\}$. The second row shows the results of multishift ESD. The third row displays results obtained with multishift ESD and with regularization (on all sources). Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices among randomly generated 3×3 matrices.

4. RESULTS

The regularization parameters α_i were set to 1000 for regularized sources, and to 0 for not regularized sources, for the benchmarks described in the following. Evaluation of different values for the α_i indicated this to be a good choice.

4.1. Results using the 3×3 mixing matrix A1

For ESD with two shifts $\Delta\mathbf{r} \in \{(0,0),(5,5)\}$ we obtained a high success rate for separation, even at high noise levels (figures 2a and 2d), but the mean reconstruction error for both datasets is relatively bad (0.55) for signal to noise ratios (SNR) below 5 dB, even though the variance is low over the whole SNR range.

Using ESD with multiple shifts,[1], the number of successful separations decreased (see figures 2b and 2d), because gradient descent sometimes converges to local minima of the cost function. Nevertheless, at high noise levels the separation is on average much better than when using only two shifts.

Now we introduce our prior knowledge about the time course of all three of the original mixtures. The gradient descent, starting with a random $\mathbf{W}$, on the cost function (equation (6)) is much more stable using this method. Each of the 15 trials at each noise level was successful. Also the reconstruction error at high noise levels is below 0.2 with a small variance and is therefor much better than the results with two shifts.

Figure 3 shows the results of the blind source separation, when only one of the sources was regularized (only $\alpha_1 \neq 0$, i.e. regularization on first column of $\mathbf{W}^{-1}$). Comparing figures 3a and 3c with figure 2 we can see that the number of successful separations is reduced and the mean error and the variance are worse than in the case with prior knowledge on all sources. On the other hand these plots still show an enhancement when compared to multishift ESD without prior knowledge.

Part of this decrease in the separation performance is due to the fact that the source for which we use prior knowledge in the regularization term is well separated, whereas the other sources are still mixed in some trials. In this case the calculation of the mean reconstruction error as introduced in equations (7) and (8) is not appropriate anymore. Instead, only the sources of interest should be included in calculating the error. Figure 4 shows a typical result for a SNR of 0 dB. We found that we can stabilize the separation performance by initializing **W** with the inverse of a matrix, which contains the assumed time course (which is part of our prior knowledge) in the first column and random noise in the others (see figures 3b and 3d).

4.2. Results using the 3×10 mixing matrix A2

Usually BSS algorithms estimate as many sources as there are mixtures. In applying these algorithms to real world data, this leads to two problems, which influence the number of mixtures one wants to observe: (1) Often the number of sources underlying the observed mixtures is not known.

[1]which are arranged in a star-like pattern with shifts from the set $\{(0,\pm r),(\pm r,0),(\pm r,\pm r)\}$, with $r \in \{0,5,10,15,20,30\}$

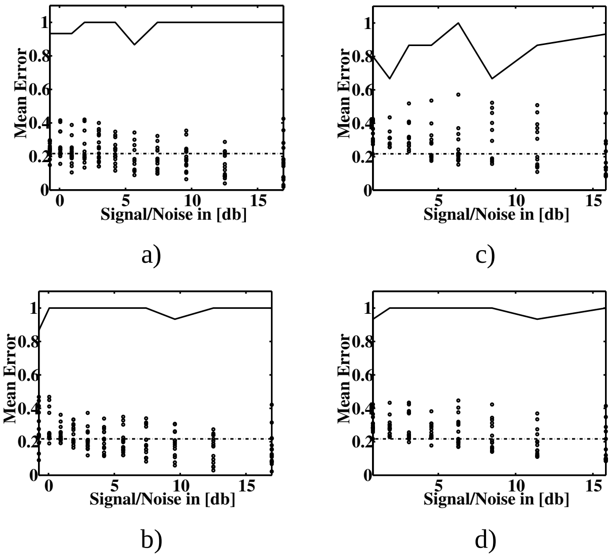


Figure 3: Mean reconstruction error as a function of the signal to noise ratio in dB (15 trials per noise level) for the mixing matrix **A2**. The left column shows the results of the separation for the smooth and the right column for the natural sources. In the first row the results with multishift ESD and the regularization term on the time course of only the first sources is displayed. A better and more stable convergence of the gradient descent can be achieve by initializing the estimate of **W** with the inverse of a matrix, that has the assumed time course in the first column and random noise in the others (bottom row). Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices among randomly generated of 3×3 matrices.

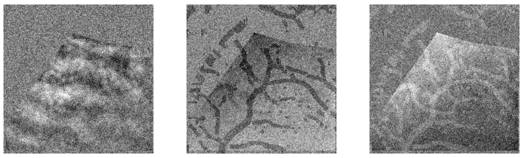
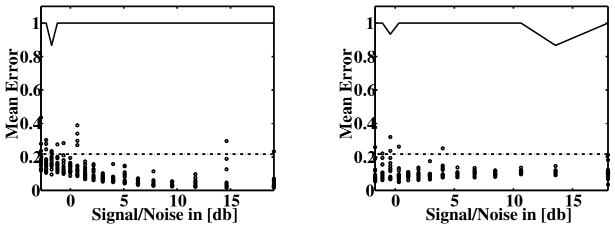


Figure 4: The three separated natural sources after application of multishift ESD and the regularization term on only the first time course of the first source at the high signal to noise ratio of 0 dB. The source of interest is well separated, whereas the other two sources are still mixed.



a) b)

Figure 5: Mean reconstruction error as a function of the signal to noise ratio in dB (15 trials per noise level). For both plots the original sources where mixed with the 3×10 matrix **A2**. For the demixing process the multishift ESD algorithm with the regularization term for the first three of the ten mixtures was used. a) shows the result for the smooth sources and b) the result for the natural sources. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3×3 matrices.

(2) One does not want to throw away information contained in the observed mixtures, even though the number of available mixtures may be higher than the estimated number of sources. To evaluate the effect of using more mixtures than there are sources, in a noisy environment, we created 10 mixtures for each of the two sets of three sources. This also gives a hint on the scaling properties of the regularized ESD algorithm, as we now have to estimate a 10×10 demixing matrix, instead of a 3×3-dimensional one. Applying the regularization term to the first three of the ten columns of $\mathbf{W}^{-1}$ we can force the underlying original sources into the first three estimated sources. The calculation of the mean reconstruction error as given in equations (7) and (8) is then applied to those first three estimated sources.

Figure 5 shows the separation result for the smooth sources (a) and the natural sources (b) down to signal to noise rations below 0 dB. It shows that the minimization of the cost function in equation (6) converges well even for the larger numbers of mixtures. The percentage of successful separations is almost 100% over the whole SNR range. Also the variance around the mean reconstruction error is low at all noise levels.

5. SUMMARY AND CONCLUSIONS

We have introduced a regularization term into the cost function of ESD with multiple shifts, which incorporates prior information about the time courses of the original sources. In this term the distance between the time course of the estimated sources and an given time course is scored.

For the ESD algorithm with only two shifts we can rewrite the cost function of equation (3) as an eigenvalue problem [5, 1]. If we use multiple shifts then much better separation results can be achieved, but the number of successful separations can decrease. This happens because the gradient descent algorithms can run into local minima. When we use the additional information about estimated time courses of the sources the gradient descent method improves in stability (see figures 2c and 2f). For the case that we use prior knowledge about the time course of only one of the sources the algorithm can be stabilized if we initialize **W** with the inverse of a matrix, which contains the assumed time course of that source in the first column and random noise in the others. In previous work [4] we used spatial lowpass filtering to improve the reconstruction error. When compared to the methods used here, we can see that a similar or smaller error can be achieved by using prior knowledge about the time course of signals. This approach avoids the issue of possibly introducing artifacts into the sources by filtering the mixtures [9].

With the regularized multishift ESD we get much better separation of the original sources, but several trials may be necessary to get a successful separation. Application of the suggested regularization method can help to improve noise robustness, and it can separate signals of interest into given sources, avoiding the permutation problem of standard BSS approaches.

6. REFERENCES

[1] I. Schießl, M. Stetter, J. E. W. Mayhew, S. Askew, N. McLoughlin, J. B. Levitt, J. S. Lund, and K. Obermayer, "Blind separation of spatial signal patterns from optical imaging records.," in *Proceedings of the 1. ICA99 Workshop, Aussois*, J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., 1999, vol. 1, pp. 179–184.

[2] H. Schöner, M. Stetter, I. Schießl, J. Mayhew, J. Lund, N. McLoughlin, and K. Obermayer, "Application of blind separation of sources to optical recording of brain activity," in *Advances in Neural Information Processing Systems NIPS 12*, S.A. Solla, T.K. Leen, and K.-R. Müller, Eds. 2000, pp. 949–955, MIT Press, In press.

[3] G. G. Blasdel and G. Salama, "Voltage-sensitive dyes reveal a modular organization in monkey striate cortex.," *Nature*, vol. 321, pp. 579–585, 1986.

[4] I. Schießl, M. Stetter, J. E. W. Mayhew, N. McLoughlin, J. S. Lund, and K. Obermayer, "Blind signal separation from optical imaging recordings with extended spatial decorrelation," *IEEE Trans. Biomed. Engin.*, p. in press, 2000.

[5] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, pp. 3634–3637, 1994.

[6] S. M. Rüger, "Stable dynamic parameter adaptation.," in *Advances in Neural Information Processing Systems.*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., vol. 8, pp. 225–231. MIT Press Cambridge, MA, 1996.

[7] K.-R. Müller, Philips P, and A. Ziehe, "Jadetd: Combining higher-order statistics and temporal information for Blind Source Separation (with noise).," in *Proceedings of the 1. ICA99 Workshop, Aussois*, J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., 1999, vol. 1, pp. 87–92.

[8] B.-U. Koehler and R. Orglmeister, "Independent component analysis using autoregressive models.," in *Proceedings of the ICA99 workshop*, J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., 1999, vol. 1, pp. 359–363.

[9] M. Stetter, T. Otto, T. Mueller, F. Sengpiel, M. Huebener, T. Bonhoeffer, and K. Obermayer, "Temporal and spatial analysis of intrinsic signals from cat visual cortex.," *Soc. Neurosci. Abstr.*, vol. 23, pp. 455, 1997.