Supplementary Information

Causality principle in reconstruction of sparse NMR spectra

Maxim Mayzel[a], Krzysztof Kazimierczuk[b], and Vladislav Yu. Orekhov[a,*]

Theory

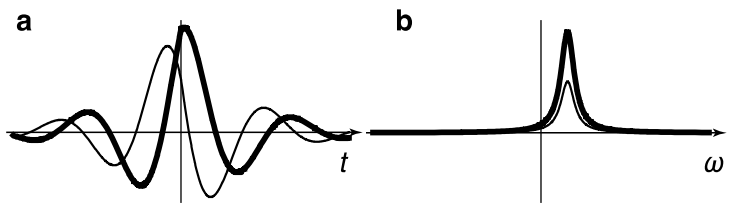


Figure S1. (**a**) The true echo time signals from Eq. 2c and (**b**) the corresponding spectrum (Eq. 3c). Small φ=0.15π is used to illustrate the effect of non-zero phase on the signal in the time and frequency domains.

True and virtual echoes

The virtual-echo time-domain presentation given by Eq. 1 should be distinguished from the traditional true-echo signal previously suggested for obtaining pure-absorption spectra [1, 2]. The former is a mathematically equivalent presentation of a regular FID, the latter requires a specialized experimental setup. Below we illustrate relation between the virtual and true echo for the specific case of a single-component NMR signal with frequency Ω, transverse relaxation rate α, and phase ϕ. Then, the FID and VE time signals are

$$S_{FID}(t) = \theta(t)\exp[- i\Omega t - \alpha t + i\phi] \quad - \infty < t < \infty \qquad (S1a)$$

$$S_{VE}(t) = \exp[- i\Omega t - \alpha|t| + i\phi\ sign(t)] \quad - \infty < t < \infty \qquad (S1b)$$

where $\theta(t)$ is the Heaviside step function. Equation S1b, which is derived using Eqs. 1 is an equivalent presentation obtained from the original FID signal (Eq. S1a) without adding or loosing information. In general case of $\phi \neq 0$, as it can be seen in Fig. 1c, $S_{VE}(t)$ has a discontinuity at the time point zero. The true-echo signal, which is observed in the spin-echo type experiments [1, 3, 4] and used in pseudo-echo signal processing [5]

$$S_{TE}(t) = \exp[- i\Omega t - \alpha|t| + i\phi] \quad - \infty < t < \infty \qquad (S1c)$$

is continuous for any phase ϕ. Fourier transform of the signals from equations S1 gives the following spectra:

$$F_{FID}(\omega) = \cos(\phi)L(\omega) + \sin(\phi)D(\omega) + i[\sin(\phi)L(\omega) - \cos(\phi)D(\omega)] \qquad (S2a)$$

$$F_{VE}(\omega) = 2\cos(\phi)L(\omega) + 2\sin(\phi)D(\omega) \qquad (S2b)$$

$$F_{TE}(\omega) = 2\cos(\phi)L(\omega) + 2i\sin(\phi)L(\omega) \qquad (S2c)$$

where $L(\omega)$ and $D(\omega)$ are absorption and dispersion parts:

$$L(\omega) = \frac{1}{\sqrt{2\pi}} \frac{\alpha}{\alpha^2 + (\Omega - \omega)^2} \quad (S3a)$$

$$D(\omega) = \frac{1}{\sqrt{2\pi}} \frac{\Omega - \omega}{\alpha^2 + (\Omega - \omega)^2} \quad (S3b)$$

Signals and corresponding spectra given by Eqs. S1-3 are illustrated in Figs. 1 and S1. Equations S2 and S3 explain differences between spectra produced from the normal FID and echo signals. The former always contain equal amount of the absorption and dispersion, whose contributions to real and imaginary parts of the spectrum depend on the signal phase. Virtual-echo spectrum has zero imaginary part. Its real part contains mixture of absorption and dispersion parts. If the phase is known, $F_{VE}(\omega)$ can be adjusted to the pure absorption mode. The method relies on prior knowledge about the phase and thus has limited use when the phase is not known, e.g. if it is the parameter to be measured in experiment or if spectrum contains signals of significantly different phases.

The true-echo spectrum is purely absorptive, with the signal power distributed between the real and imaginary parts. Both the true and virtual echo approaches can be used to produce dark, pure absorption spectrum and thus provide basis for efficient sparsifying transforms. The conventional MRI signal in the *k*-space corresponds to the true-echo, which contributes to success of the compressed sensing techniques in that field. However, to our knowledge, the true-echo has not been used so far in relation to the NMR spectra reconstruction from NUS data. A signal representation similar to VE was suggested for spectra reconstruction from projections [6] and was also used for NUS spectra reconstruction with a new iterated maps algorithm [7]. However there, the VE was entangled into the method-specific algorithms and general implications of the VE presentations for NUS spectra reconstruction was not discussed.

Digital considerations

In practical NMR experiment, N signal points are sampled at regular time intervals Δ over a period of time $t_{max} = \Delta(N\text{-}1)$. Considering that the discrete Fourier transform (DFT) assumes that the signal is periodic with the period t_{max} and starts at the time point zero, equations for the virtual echo presentation of the digitized signal are:

$$S_{VE} = [\, (S_0 + S_0^*)/2,\ S_1,\ldots,S_{N-1},\ 0,\ S_{N-1}^*,\ S_{N-2}^*,\ \ldots,\ S_1^* \,] \quad (S4a)$$

$$S_{VE} = [\, S_0,\ S_1,\ldots,S_{N-1},\ S_{N-1}^*,\ S_{N-2}^*,\ \ldots,\ S_1^*,\ S_0^* \,] \quad (S4b)$$

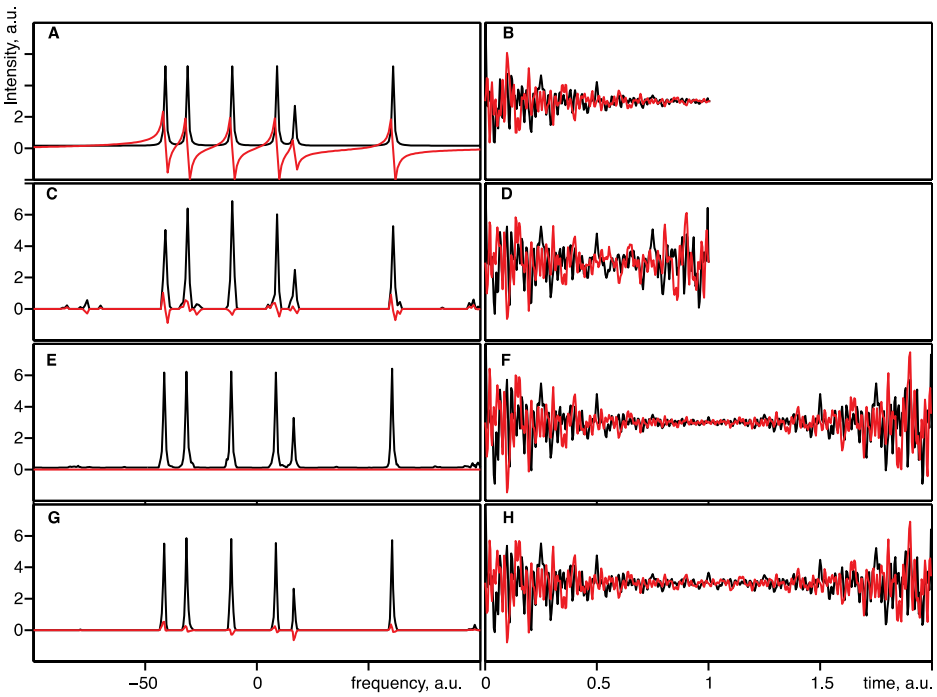
where we consider two cases for position of the first measured data point: (Eq. S4a) at time point t=0, i.e. when all frequency components of the FID have the same phase, and (Eq. S4b) at the half dwell time, $t = \Delta/2$. Positions of the first data point, which correspond to the first order phase correction in the spectrum other than $k\pi, k \in N$ are not compatible with the VE presentation. Eq. S4a can be seen as a sum of two time domain signals: right-zero-padded up to 2N points FID:

$$S_+ = [\, S_0/2,\ S_1,\ldots,S_{N-1},\ 0,\ldots,0 \,] \quad (S5a)$$

and the complex conjugate of the time reversed FID, which is also left-zero-padded and cyclically shifted by one point:

$$S_-^r = [\, S_0^*/2,\ 0,\ldots,0,\ S_{N-1}^*,\ S_{N-2}^*,\ \ldots,\ S_1^* \,] \quad (S5b)$$

In the frequency domain, signals in Eqs. S5a and S5b have the same absorption parts. The dispersion parts, however, have the opposite signs and thus cancel out in the VE signal in Eq. S4.
If zero order phase ϕ of the signal is known, it should be applied to the time-domain FID signal prior to the conversion to the VE. Then, the spectrum reconstruction from the time signal in Eq. S4a consists of the pure



absorption real part and the zero imaginary part. With corrected zero order phase, spectrum of the signal in Eq. S4b contains only absorption contribution, which is however distributed between the real and imaginary parts. As in the case of the true-echo, this does not affect darkness of the spectrum and, if needed, the pure absorption spectrum with zero imaginary part can be obtained by the linear phase correction of 180^0 in the frequency domain.

Figure S2. Effect of FID extrapolation and VE presentation on IST spectrum reconstruction from 20% non-uniformly sampled data. (**a**) Synthetic reference 1D spectrum consisting of six signals and (**b**) the corresponding complete FID. (**c,e,g**) reconstructed spectra. (**d,f,h**) reconstructed time domain signals. (**c,d**) IST on NUS FID without extrapolation. (**e,f**) IST on NUS FID in VE presentation. (**g,h**) IST on NUS FID with extrapolation.

Relation to zero filling and extrapolation

A similar to the VE signal presentation yet without the zero-padding in Eqs. S5 was suggested by Nagayama [8] for elimination of the dispersion signals. As it was later pointed out by Szyperski and co-workers [2], Nagayama's approach is not suitable for producing quantitative spectra. We must stress that the zero padding (also called zero-filling) of the time domain signals in Eqs. S5 serves for ensuring causality of the NMR signal in digital analysis, which in the traditional Fourier based NMR processing leads to enhancement of the resolution and notably sensitivity in the spectra [9]. However, extrapolation of incompletely decayed signal by zeroes results in the sinc-type distortions of the spectral line shapes. Thus, in case of non-uniform sampling zeroes in Eqs. S5 are not used explicitly but considered as missing data. Consequently, NUS VE signal (Eq. S4, Fig S2f) has no explicit zeroes. Comparison of panels c and g in Figure S2 illustrates that, similar to the traditional Fourier signal processing, the extrapolation of the FID significantly improves quality of the NUS spectra reconstruction. Although never explicitly discussed previously, extrapolation is usually implemented in the NMR CS algorithms. Also in this work, for comparison with the VE, the FID is always extrapolated two times in the course of the CS and SIFT reconstructions.

Since, CS finds the sparsest spectrum in the frequency domain, it tends to suppress the relatively low intensity dispersion tails of the signals (Fig. S2 c, g), which inevitably leads to a signal resembling the VE (Fig. S2 d, h). For the non-extrapolated signal, CS algorithm significantly distorts the time domain signal (Fig S2d) in comparison with the reference FID (Fig S2b) and, correspondingly, quality of the reconstructed spectrum (Fig S2c) is poor. When the time signal is extrapolated, the missing data in the extrapolated part allows for accommodation of the time-reversed signal (Fig S2c), which is enforced by the CS sparseness constraint. Thus, the correct FID is reconstructed and the spectrum is fine. In this context, the VE method (Fig. S2 e, f) fits naturally to CS behaviour – it uses the experimental points where otherwise the reconstruction would take place. The number of unknowns is thus reduced and the reconstruction performs better.

Generalization for N-dimensions

The virtual-echo transformation defined by Eq. 5 can be easily generalized for a spectrum of any dimensionality. For example, for every time point (t_1,t_2) of a two-dimensional signal $(0 \le t_1 < t1_{max},\ 0 \le t_2 < t2_{max})$, States method of NMR signal detection gives four real measurements, which are essentially products of the form R_1R_2, R_1I_2, I_1R_2, and I_1I_2. From these we can build four complex matrices:

$$S_{\pm\pm}(t_1,t_2) = (R_1 \pm iI_1)(R_2 \pm iI_2) \qquad (S6)$$

For example, complex matrix $S_{++}(t_1,t_2)$ is obtained as

$$S_{++}(t_1,t_2) = (R_1 + iI_1)(R_2 + iI_2) = R_1R_2 - I_1I_2 + i[I_1R_2 + R_1I_2]$$

Then, we double the sizes of the matrices in Eq. S6 by zero padding and perform the time-reverse operations for the individual dimensions similar to Eq. S5. Finally, after summation of the four resulting matrices we obtain a complex matrix of the 2D virtual-echo signal $(0 \le t_1 < 2\,t1_{max},\ 0 \le t_2 < 2\,t2_{max})$:

$$S_{VE}(t_1,t_2) = \begin{bmatrix} S_{++}(t_1,t_2) & S_{+-}(t_1,\ t2_{max} - t_2) \\ S_{-+}(t1_{max} - t_1,t_2) & S_{--}(t1_{max} - t_1,\ t2_{max} - t_2) \end{bmatrix} \qquad (S7)$$

The two-dimensional Fourier transform or CS reconstruction of $S_{VE}(t_1,t_2)$ signal produces complex 2D spectrum with zero imaginary part. When phases of the signal are known *a priori* and corrected prior to the reconstruction, the spectrum is obtained in the pure absorption mode. Otherwise, the spectrum may contain non-zero dispersion part, which lessen the spectral sparsity, but still allows successful reconstruction, although with more experimental data.

As the final remark, we note that while the original FID for a multidimensional spectrum is hyper-complex, the VE signal presentation is always complex and, thus, directly amenable for processing using N-dimensional DFT, CS, and other algorithms dealing with complex signals. The difference may be significant in practice for NUS data. Out of four matrices in Eq. S6, any two, e.g. $S_{++}(t_1,t_2)$ and $S_{\pm-}(t_1,t_2)$, are sufficient to present all information contained in the original hyper-complex data set. In the traditional approach, spectral reconstructions are obtained separately for these two complex time domain matrices and the final spectrum is obtained as a sum of the obtained results. Due to inherent nonlinearity of the spectral reconstructions from NUS data, imperfections in the two subparts may manifest themselves as artefacts, e.g. like distortions of the line shapes (visible in Fig. 2c). It is noteworthy that regardless of the method used for separation of the hyper-complex signal to two complex matrices, each of the complex spectra subparts is naturally less sparse than the full hypercomplex representation and thus requires more data for successful reconstruction. The effect scales up as number of subparts increases with more spectral dimensions and thus VE is likely to improve significantly high-dimensional techniques.

Experimental

1.2 mM sample of human alpha-synuclein (SWISSPROT accession number P37840) in sodium phosphate buffer 20 mM, pH 6.5, was used to obtain fully sampled 1H-15N HSQC spectrum at 15 °C on 800 MHz Agilent DDR2 spectrometer equipped with a cryogenically cooled HCN probe. The spectrum 512 complex points were acquired for the 15N dimension corresponding to the acquisition time of 177 ms. The 3D NUS HNCO spectrum of 1 mM human ubiquitin was acquired at 25 °C on 600 MHz Varian UNITY Inova spectrometer with a cold probe using 6.1% NUS out of the Nyquist grid of 120 x 79 points (43 ms and 28 ms) for the 13C and 15N spectral dimensions, respectively. For the alpha-synuclein spectrum, the NUS time-domain data were produced by selecting measurements from the fully sampled spectra in accordance with NUS schedules generated by the program nussampler from the MDDNMR software package [10]. For the ubiquitin spectrum, different NUS data were produced by random selection from the measured data points. SIFT algorithm was implemented in MATLAB as described in the original paper [11]. Unless specifically indicated otherwise, signal support in the SIFT calculations included area ±26 Hz 1H, ±10 Hz 15N around every peak in the spectrum. CS-IRLS and CS-IST calculations with and without VE were performed using MDDNMR software [12] . For measuring intensities in the spectra reconstructed from VE with the uncorrected zero order phase (inset in Fig. 2e), the spectrum was phased after the reconstruction in the frequency domain using the Hilbert transform. For SIFT and CS-IRLS/IST processing of the traditional FID, i.e. without VE, the time signal was extrapolated to twice the size in all indirect dimensions as discussed in the Theory section. The extrapolated points were considered as missing measurements in the same way as other missing data in the NUS time domain signal.

1. A. Bax, A. F. Mehkkopf and J. Smidt, *Journal of Magnetic Resonance*, 1979, **35**, 373-377.
2. A. Ghosh, Y. Wu, Y. He and T. Szyperski, *J. Magn. Reson.*, 2011, **213**, 46-57.
3. E. L. Hahn, *Phys. Rev*, 1950, **80**, 580–594.
4. Y. Wu, A. Ghosh and T. Szyperski, *Angew. Chem. Int. Ed.*, 2009, **48**, 1479-1483.
5. A. Bax, R. Freeman and G. A. Morris, *Journal of Magnetic Resonance (1969)*, 1981, **43**, 333-338.
6. B. E. Coggins and P. Zhou, *Journal of Magnetic Resonance*, 2006, **182**, 84-95.
7. M. A. Frey, Z. M. Sethna, G. A. Manley, S. Sengupta, K. W. Zilm, J. P. Loria and S. E. Barrett, *Journal of Magnetic Resonance*, 2013, **237**, 100-109.
8. K. Nagayama, *Journal of Magnetic Resonance (1969)*, 1986, **66**, 240-249.
9. E. Bartholdi and R. R. Ernst, *Journal of Magnetic Resonance (1969)*, 1973, **11**, 9-19.
10. V. Y. Orekhov and V. A. Jaravine, *Prog Nucl Mag Res Sp*, 2011, **59**, 271-292.
11. Y. Matsuki, M. T. Eddy and J. Herzfeld, *J Am Chem Soc*, 2009, **131**, 4648-4656.
12. K. Kazimierczuk and V. Y. Orekhov, *Journal of Magnetic Resonance*, 2012, **223**, 1-10.

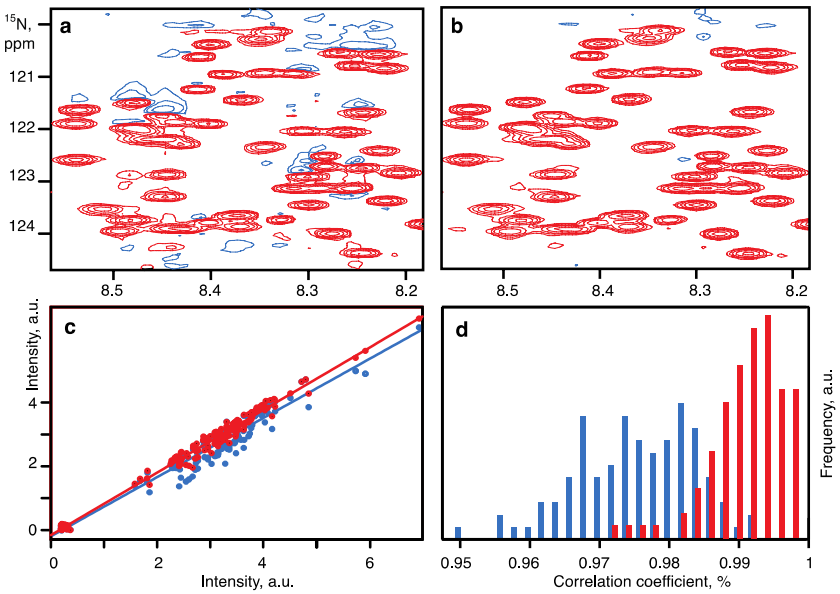


Figure S3. Comparison of CS IST spectral reconstruction of 30% NUS 2D ^{1}H-^{15}N HSQC of alpha-synuclein obtained using time-domain signal in traditional FID (**a**) and VE (**b**) presentation. (**c**) Accuracy of the peak intensities in spectra (**a**) and (**b**) are shown with (*blue*) and (*red*), respectively. Intensities of the peaks in the reconstructed spectra are shown against the corresponding intensities in the fully sampled reference spectrum. Dashed lines correspond to linear regression fit of VE (*red*) and regular FID (*blue*) data. Correlation coefficients for the peak intensities are 0.99 and 0.97 for the IST reconstruction with and without VE, respectively. Histogram (**d**) shows distribution of the correlation coefficients between signal intensities measured in the reference spectrum and the spectra reconstructed with VE (*red*) and traditional FID (*blue*) for 100 resampling trials

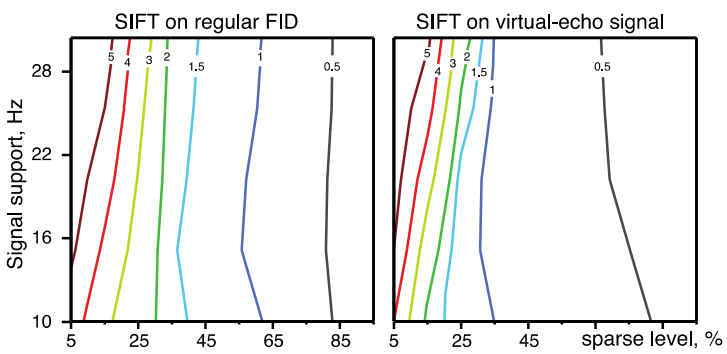


Figure S4. SIFT signal reconstruction quality scores for ^{1}H-^{15}N HSQC spectrum of alpha-synuclein versus sampling level and size of the signal support mask for the ^{15}N spectral dimension. The score is calculated as an RMSD of the difference between the reference spectrum (fully sampled, processed with DFT) and the SIFT reconstruction applied to regular FID (*left*) and to VE signal (*right*). The RMSD is calculated over all signal regions (manually verified peak maximum ± 20 Hz ^{1}H and ±10 Hz ^{15}N) and normalized to the noise level in the full reference spectrum. For every spectrum reconstruction, 10 resampling trials with different NUS schedules were performed and the reported value corresponds to the mean score. For any given sampling level, the VE shows consistently lower (i.e. better) RMSD scores and, thus better reproduces the peak intensities and line shapes.

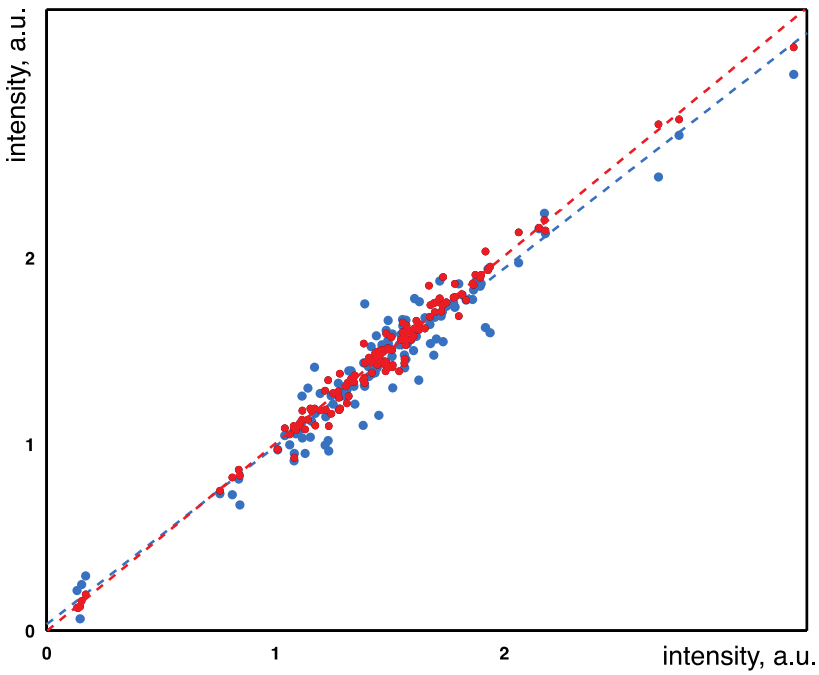


Figure S5. Accuracy of the peak intensities in the 2D ^{1}H-^{15}N HSQC spectra of alpha-synuclein shown in Figures 2a and 2b in the main text. 15% NUS spectrum was processed with SIFT applied to the VE signal (*red*) or to regular FID (*blue*); in both cases size of the signal support region in ^{15}N dimension was set to ±10 Hz around the peak maxima. Intensities of the peaks in the reconstructed spectra are shown against the corresponding intensities in the fully sampled reference spectrum. Dashed lines correspond to linear regression fit of VE (*red*) and regular FID (*blue*) data. Correlation coefficients for the peak intensities are 0.99 and 0.96 for the SIFT reconstruction with and without VE, respectively.