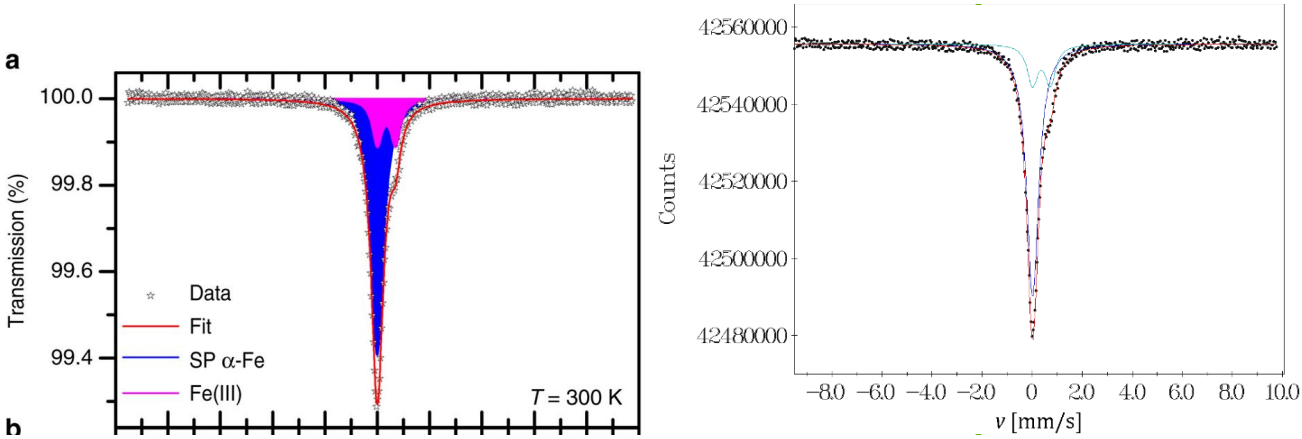
Filtrace dat a obrázky v Nature Communications

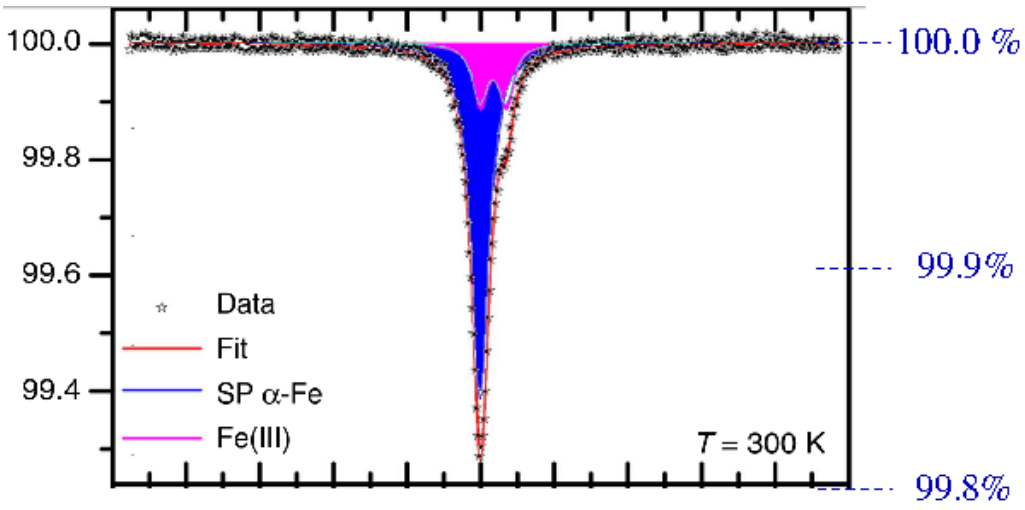
Tomáš Opatrný

18. prosince 2019

Obrázek Fig. 2a ze článku [1] (nikoliv pouze ze supplementu) vychází z dat, která poskytl děkanovi korespondenční autor ve formě obrázku (viz obr. 1 a 2). Má se jednat o tzv. „filtrovaná data“, přičemž filtrace je popsána ve článku [2].



Obrázek 1: Levý panel: Fig. 2a ze článku [1], pravý panel: obrázek poskytnutý korespondenčním autorem po žádosti o moessbauerovská data ke článku [1]. Poloha datových bodů je identická.



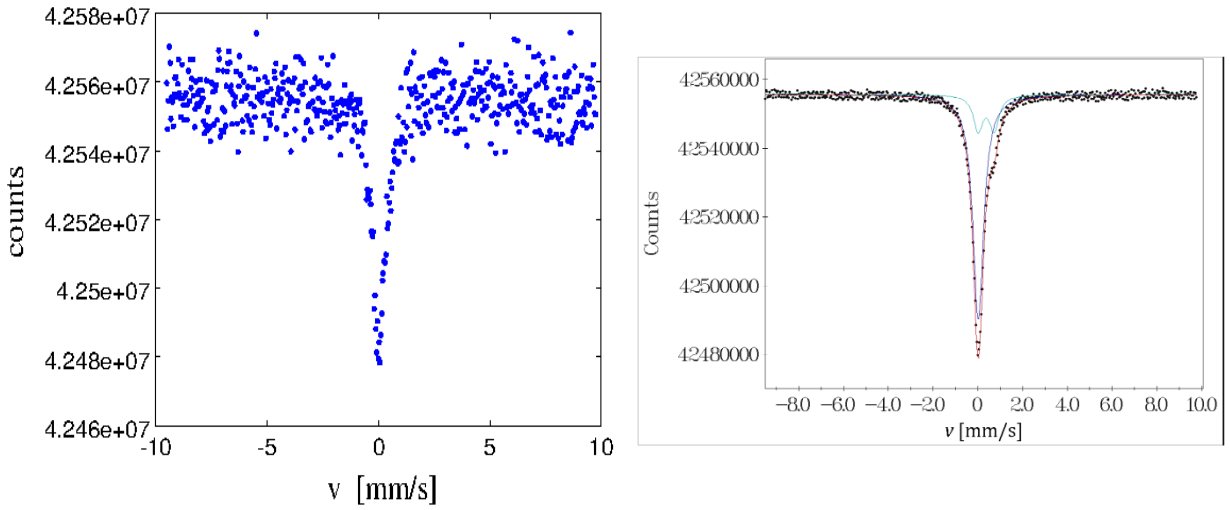
Obrázek 2: Překryv obou předchozích obrázků – datové body se přesně kryjí, byť mají odlišné procentní škály. V publikovaném článku je efekt cca 0.7 % kdežto dodaným datům odpovídá efekt 0.18%. Obrázek v článku má tedy efekt navýšený na cca čtyřnásobek původní hodnoty.

Na tomto obrázku je (krom odlišné procentní škály reportované ve článku a škály odpovídající datům) zvláště nápadný velmi nízký šum a velmi těsné přimykání datových bodů k fitovací křivce. Toto lze kvantifikovat následujícím způsobem. Fitovací křivka velmi přesně odpovídá Lorentzovým

křivkám

$$f(v) = A - \frac{B_1}{1+\left(\frac{v-v_1}{\Delta v1}\right)^2} - B_2\left(\frac{1}{1+\left(\frac{v-v_2-\frac{\delta v_2}{2}}{\Delta v2}\right)^2} + \frac{1}{1+\left(\frac{v-v_2+\frac{\delta v_2}{2}}{\Delta v2}\right)^2}\right) \tag{1}$$

s následujícími parametry: $A = 4.255563 \times 10^7$ (pozadí), $B_1 = 6.40 \times 10^4$ (amplituda singletu), $B_2 = 1.0171 \times 10^4$ (amplituda dubletu), $v_1 = 0.00765$ mm/s (poloha centra singletu), $v_2 = 0.36271$ mm/s (poloha centra dubletu), $\delta v_2 = 0.70954$ mm/s (vzdálenost píků dubletu), $\Delta v_1 = 0.29195$ mm/s (pološířka singletu), $\Delta v_2 = 0.25577$ mm/s (pološířka píku dubletu). Polohu datových bodů lze s jistou mírou přesnosti odečíst z grafu (využití aplikace `https://apps.automeris.io/wpd/`). Po odečtení fitovací křivky dostaneme šumová data. Ta mají směrodatnou odchylku $\sigma_1 \approx 860$. To lze srovnat se směrodatnou odchylkou σ_0 poissonovského šumu pozadí, která by měla nefiltrovaná data tedy $\sigma_0 = \sqrt{4.255 \times 10^7} \approx 6500$. Poměr těchto směrodatných odchylek je $\sigma_0/\sigma_1 \approx 7.5$, **filtrace tedy měla způsobit více než sedminásobný pokles směrodatné odchylky šumu**. Tato změna je ukázána na obr. 3. Jak ukazujeme níže, **toto není možné při zachování tvaru spektra**.



Obrázek 3: Levý panel: simulace měření se signální funkcí (1) a stejným počtem countů jako na obr. 1. Pravý panel: obrázek dodaný korespondenčním autorem. Oba panely mají stejné škály na horizontální i vertikální ose. Filtrace by měla vést od dat chovajících se jako na levém panelu k podobě na pravém panelu.

Filtrace popsaná ve článku [2] je založena na odstraňování komponent Fourierova obrazu spektra. Odstraňují se jednak komponenty s nejvyšší frekvencí (tzv. „high-cut filter“) a jednak ty s nejmenší absolutní hodnotou (tzv. „statistical filter“). Fourierovské komponenty šumu se spočítají jako

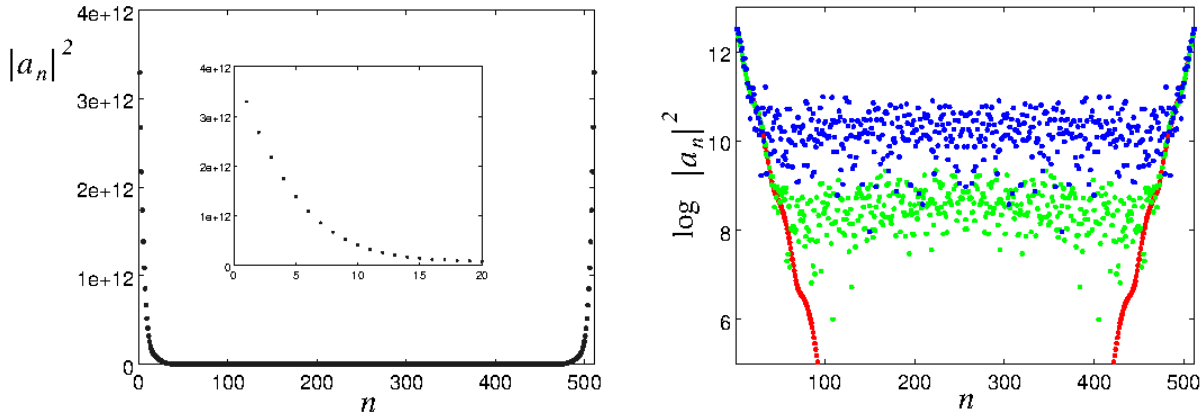
$$a_n = \sum_{k=0}^{N-1} f_k e^{i\frac{2\pi kn}{N}}, \tag{2}$$

kde N je počet naměřených hodnot ($N = 512$ pro moessbauerovská data) a f_k jsou šumové komponenty splňující (pro původní data)

$$\langle f_k f_l \rangle = \delta_{k,l}\sigma_0^2, \tag{3}$$

kde $\delta_{k,l}$ je Kroneckerovo delta. Střední hodnota kvadrátu každé fourierovské komponenty šumu pak je

$$\langle |a_n|^2 \rangle = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \langle f_k f_l \rangle e^{i\frac{2\pi(k-l)n}{N}} = N\sigma_0^2. \tag{4}$$



Obrázek 4: Levý panel: fourierovské spektrum signální funkce (1). Pravý panel: logaritmus fourierovských komponent signální funkce (červené body), logaritmus fourierovských komponent simulovaných detekcí s poissonovskou statistikou (modré body) a datových hodnot z obr. 1 (zelené body).

Pokud provedeme filtraci (vynulujeme některé fourierovské komponenty) a poté zpětnou Fourierovu transformaci, budou filtrovaná šumová data rovna

$$\tilde{f}_k = \frac{1}{N} \sum_{n \in S} e^{-i\frac{2\pi kn}{N}} a_n, \tag{5}$$

kde S je množina indexů, pro které fourierovské komponenty nebyly vynulovány. Předpokládejme, že počet prvků této množiny je N'. Variance (kvadrát směrodatné odchylky) šumových dat pak je

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{N} \sum_{k=0}^{N-1} |\tilde{f}_k|^2 = \frac{1}{N^3} \sum_{k=0}^{N-1} \sum_{n \in S} \sum_{m \in S} a_n^* a_m e^{i\frac{2\pi k(n-m)}{N}} \\ &= \frac{1}{N^3} \sum_{n \in S} \sum_{m \in S} a_n^* a_m N \delta_{m,n} = \frac{1}{N^2} \sum_{n \in S} |a_n|^2. \end{aligned} \tag{6}$$

Pokud v posledním výrazu nahradíme $|a_n|^2$ střední hodnotou dle (4), dostaneme

$$\tilde{\sigma}^2 \approx \frac{N'}{N} \sigma_0^2. \tag{7}$$

Tato aproximace je dobrá v případě, kdy množina S obsahuje komponenty napříč všemi velikostmi, tedy jako v případě filtrace metodou „high cut filter“. V případě filtrace metodou „statistical filter“, kdy se odstraňují komponenty s nejmenší absolutní hodnotou, obsahuje množina S komponenty s větší velikostí a výsledná variance šumových dat je větší, tedy

$$\tilde{\sigma}^2 > \frac{N'}{N} \sigma_0^2. \tag{8}$$

To znamená, že pokud má být směrodatná odchylka zredukována v poměru $\sigma_1/\sigma_0 \approx 1/7.5$, musí být vynulováno velké množství fourierovských komponent tak, že počet zbylých komponent je

$$N' \lesssim \left(\frac{\sigma_1}{\sigma_0}\right)^2 N \approx 0.02\ N \approx 10. \tag{9}$$

Zredukovat šum na takovou velikost, jaká odpovídá obrázku 1, tedy znamená ponechat z původního Fourierova obrazu jen cca 10 komponent. To ale znamená, že zpětnou Fourierovou transfomací se z takto malého počtu komponent získají data se silnými korelacemi – tedy namísto nezávisle fluktuujících

bodů budou mít tvar křivek oscilujících na frekvencích zbylých komponent. Kromě toho se vynulování většiny komponent spektra projeví i jako zásah do tvaru původní signální funkce.

To je vidět na obrázcích 4, 5 a 6. Na obr. 4 je fourierovské spektrum signální funkce (1) (kvadráty absolutních hodnot fourierovských komponent) a dat získaných jednak simulací experimentu (k fitovací křivce byl přičten šum se směrodatnou odchylkou σ_0, modré body na pravém panelu) a jednak dat odečtených z obrázku 1 (zelené body). Je vidět, že spektrum signální funkce má významně velké hodnoty v rozmezí n od 0 do cca 20 a podobně od cca 490 do 512, tedy cca 40 fourierovských komponent hraje důležitou roli.

Na obr. 5 je pak vidět výsledek filtrací metodou „high-cut filter“ pro různý počet ponechaných komponent N'. Jak je vidět, s ubývajícím počtem komponent se ve filtrovaných datech objevují korelace projevující se jako oscilace provázející základní tvar signální funkce. Ty pocházejí nejprve z šumových komponent a při poklesu N' pod cca 40 také z deformace původní signální funkce.

Pro filtraci metodou „statistical filter“ jsou výsledky na obr. 6. Vzhledem k tomu, že se odstraňují komponenty s nejmenší velikostí, zůstávají ve spektru hodnoty s vyšším příspěvkem šumu.

Podobné výsledky lze získat i při použití filtrovacích metod využitých v systému MossWinn [3] - tedy např. gaussovským filtrem, filtrací Fermi-Diracovou funkcí, lineární funkcí, atd.

Závěr:

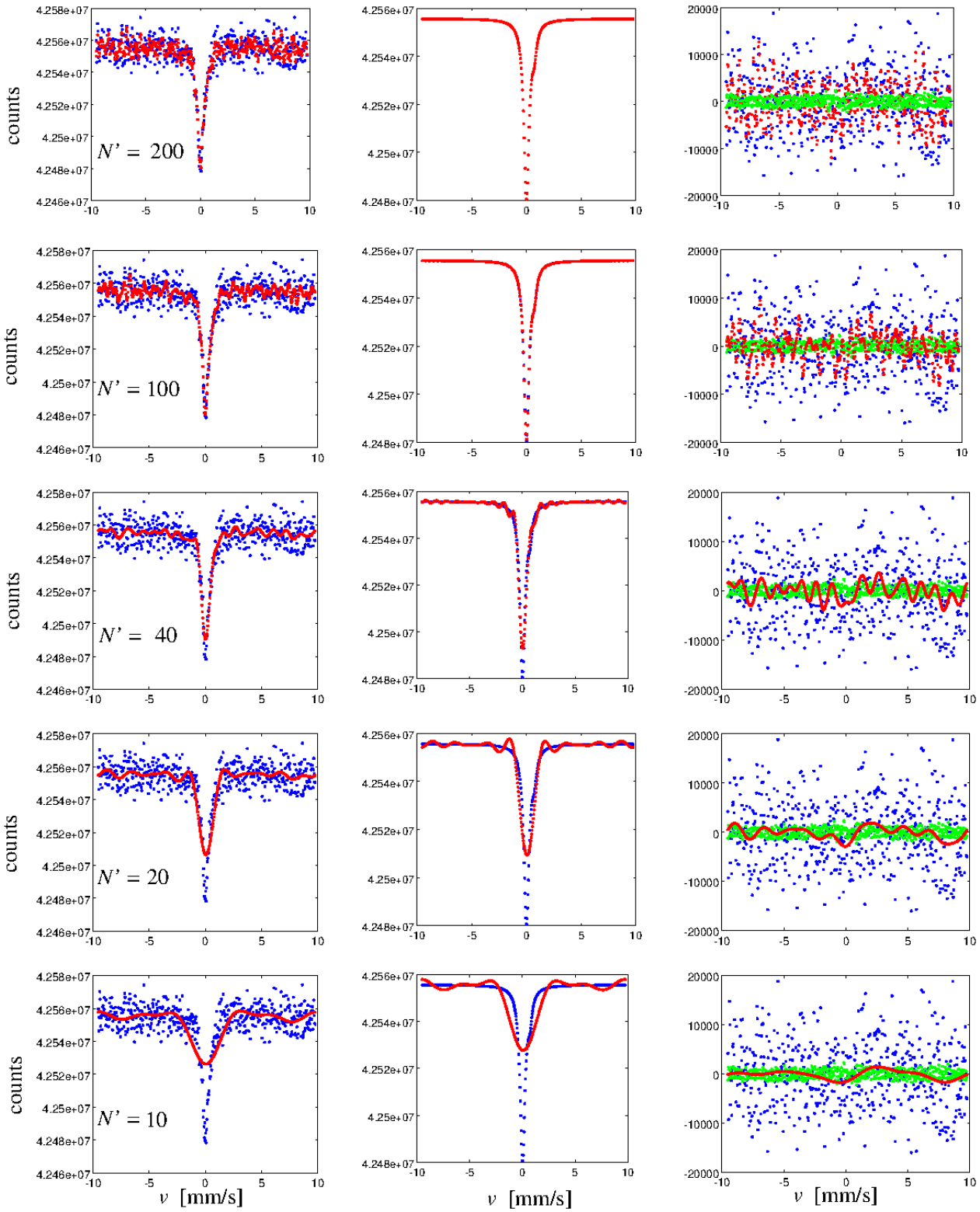
Uvedené výsledky ukazují, že data prezentovaná na obr. 1 zřejmě nejsou filtrovaná data z experimentu (krom toho, že ve článku byl efekt „navýšen“ na čtyřnásobek). Tomu, že nejde o filtraci odpovídající postupu v [2], svědčí i to, že spektrum má v celé oblasti nenulové hodnoty, pouze nižší než u spektra z dat s poissonovským šumem o odpovídající velikosti (viz obr. 4). Obrázek 1 (a tedy i Fig. 2a ze článku [1]) byl tedy nejspíš vytvořen simulací šumu kolem zvolených lorentzovských křivek.

Reference

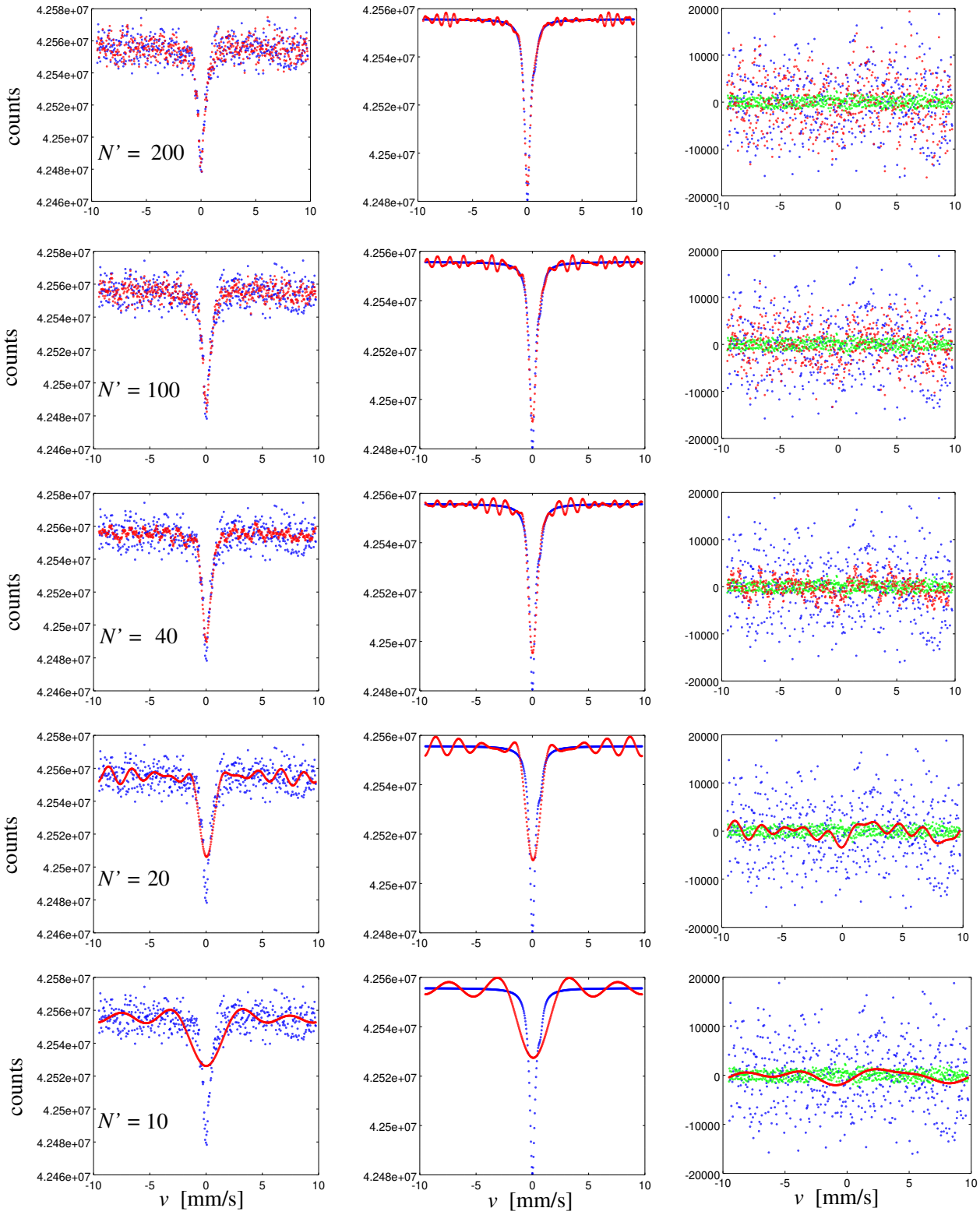
[1] Tuček et al., *Air-stable superparamagnetic metal nanoparticles entrapped in graphene oxide matrix.* Nature Communications 7:12879—DOI: 10.1038/ncomms12879 (2016).

[2] R. Procházka et al, *Statistical analysis and digital processing of the Mössbauer spectra.* Meas. Sci. Technol. 21 (2010) 025107.

[3] Z. Klencsár, *MossWinn 4.0i Manual, revision 2019.02.03*, `http://www.mosswinn.hu/downloads/mosswinn.pdf`.



Obrázek 5: Filtrace s $N' = 200$, $N' = 100$, $N' = 400$, $N' = 20$ a $N' = 10$, tzv. „high-cut filter“. Levý sloupec: původní simulovaná data (modrá) a data po filtraci (červená), prostřední sloupec: signální funkce původní (modrá) a filtrovaná (červená - uvedeno pro ilustraci, aby bylo viditelné zvlnění, které filtrace způsobí), pravý sloupec: šum původní (modrá) a po filtraci (červená). Pro srovnání zelené body ukazují šumové hodnoty odpovídající článku [1]. Výsledná směrodatná odchylka šumu je $\tilde{\sigma}_{200} \approx 4100$, $\tilde{\sigma}_{100} \approx 3000$, $\tilde{\sigma}_{40} \approx 1770$, $\tilde{\sigma}_{20} \approx 1200$ a $\tilde{\sigma}_{10} \approx 860$.



Obrázek 6: Filtrace s $N' = 200$, $N' = 100$, $N' = 400$, $N' = 20$ a $N' = 10$, tzv. „statistical filter“. Levý sloupec: původní simulovaná data (modrá) a data po filtraci (červená), prostřední sloupec: signální funkce původní (modrá) a filtrovaná (červená), pravý sloupec: šum původní (modrá) a po filtraci (červená). Pro srovnání zelené body ukazují šumové hodnoty odpovídající článku [1]. Výsledná směrodatná odchylka šumu je $\tilde{\sigma}_{200} \approx 5400$, $\tilde{\sigma}_{100} \approx 4100$, $\tilde{\sigma}_{40} \approx 2040$, $\tilde{\sigma}_{20} \approx 1310$ a $\tilde{\sigma}_{10} \approx 910$.