

1 Chemical sparsity in Bayesian receptor models for aerosol 2 source apportionment

3 Marta Via¹, Jure Demšar², Yufang Hao³, Manousos Manousakas^{3,4}, Anton Rusanen⁵, Jianhui
4 Jiang⁶, Stuart K. Grange⁷, Jean-Luc Jaffrezo⁸, Vy Ngoc Thuy Dinh⁸, Gaëlle Uzu⁸, Griša
5 Močnik¹, and Kaspar R. Daellenbach³

6
7 ¹Center for Atmospheric Research, University of Nova Gorica, Ajdovščina 5270, Slovenia

8 ²Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia

9 ³Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, 5232 Villigen PSI, Switzerland

10 ⁴Environmental Radioactivity Aerosol Tech. for Atmospheric Climate Impacts, INRaSTES, National Centre of
11 Scientific Research “Demokritos”, Ag. Paraskevi, 15310, Greece

12 ⁵Atmospheric Composition Research, Finnish Meteorological Institute, 00101 Helsinki, Finland

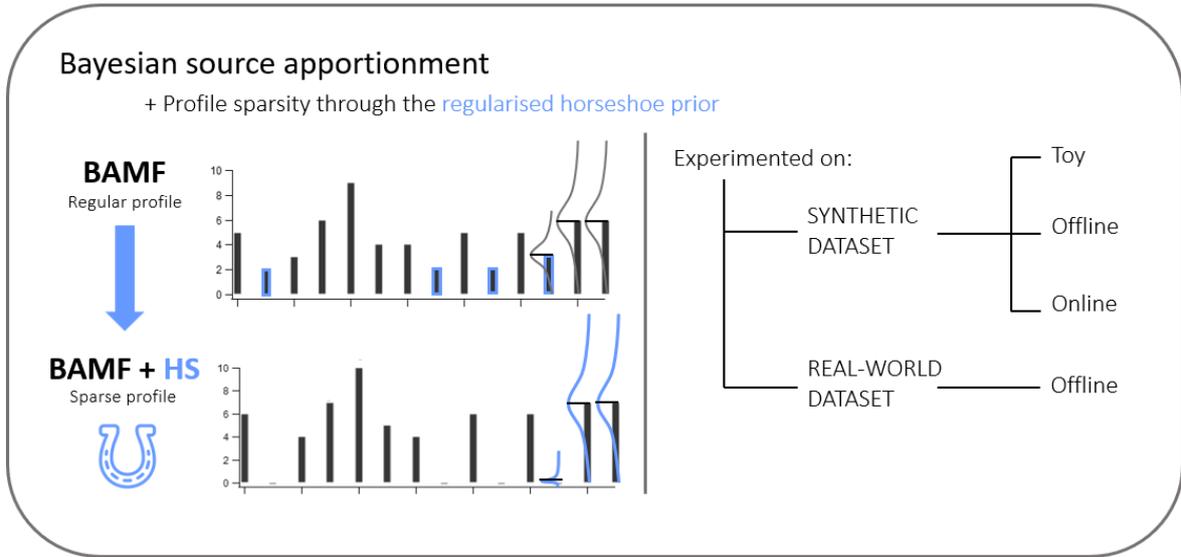
13 ⁶School of Ecological and Environmental Sciences, East China Normal University, 200241, Shanghai, China

14 ⁷Climate and Environmental Physics, Physics Institute, University of Bern, Sidlerstrasse 5, 3012 Bern,
15 Switzerland

16 ⁸University of Grenoble Alpes, CNRS, INRAE, IRD, Grenoble INP, IGE, Grenoble 38000, France

17
18 *Correspondence to:* Marta Via (marta.viagonzalez@ung.si) and Kaspar R. Daellenbach
19 (kaspar.daellenbach@psi.ch)

20 **Abstract.** Aerosol source apportionment is a key tool for understanding the origins of atmospheric particulate
21 matter and for guiding effective air quality management strategies. However, source apportionment techniques
22 still struggle to properly separate highly correlated sources without relying on restrictive *a priori* information,
23 possibly skewing the solution and adding subjective operator input, with varying degrees of benefit. This study
24 introduces sparsity into the Bayesian Autocorrelated Matrix Factorisation (BAMF) model with the aim of
25 removing non-essential species contribution in the unconstrained profiles, which is expected to improve the
26 separation of factors compared to BAMF. The regularised horseshoe prior (HS) has been added to BAMF
27 (BAMF+HS) to promote composition matrix \mathbf{F} sparsity, shrinking low-signal contributions to the solutions.
28 BAMF+HS was evaluated using three synthetic datasets designed to reflect increasing levels of data complexity
29 (Toy, representing a highly simplified dataset; Offline, representing a filter dataset; and Online, representing an
30 Aerosol Chemical Speciation Monitor (ACSM)-like dataset), and a real-world multi-site filter dataset. The results
31 demonstrate that BAMF+HS effectively enforces sparsity in offline datasets and that this improves accuracy in
32 reconstructing source profiles and time series compared to BAMF and Positive Matrix Factorisation (PMF).
33 However, its application to higher-complexity ACSM datasets revealed sensitivity to sampling instability
34 hindering sparsification. With that, even though sparsity was not achieved, the quality of the BAMF+HS solution
35 metrics were not deprecated compared to BAMF. Overall, this work underscores the value of incorporating profile
36 sparsity as a solution property in Bayesian source apportionment, and positions BAMF+HS as a promising model
37 for source apportionment.



39

40 1. Introduction

41 Particulate Matter (PM) adversely affects human health through both short and sustained exposures (Pope and
 42 Dockerty, 1999, Yang et al. 2019). The observed relationship between decreasing PM concentrations and
 43 increased life expectancy (Keuken et al. 2011; Zheng et al. 2022) highlights the importance of developing
 44 mitigation plans grounded in detailed knowledge of PM sources composition and concentrations. Moreover,
 45 because some proxies for aerosol toxicity, among them oxidative potential, are highly dependent on its sources
 46 (Daellenbach et al. 2020), implementing source-specific mitigation measures contributes to more quantitative and
 47 efficient abatement and a more effective protection of the population.

48

49 Source apportionment is the process of identifying and quantifying sources by using information about their
 50 chemical composition, and is commonly conducted through receptor models (RMs) which differentiate PM
 51 sources according to the distinctness of their chemical composition and time series characteristics. The most
 52 widely used RM is the Positive Matrix Factorisation model (PMF, Paatero and Tapper, 1994), which deconvolutes
 53 the input chemical composition into the product of composition and time series matrices (\mathbf{F} and \mathbf{G} , respectively),
 54 and minimises the residuals of the fit through the weighted least squares loss. The factorisation equation, hence,
 55 is written as

$$56 \quad \mathbf{X} = \mathbf{G} \cdot \mathbf{F} + \mathbf{E}, \quad (1)$$

57 where \mathbf{X} is the input matrix, a $n \cdot m$ matrix of n timepoints and m species, which is decomposed into \mathbf{G} and \mathbf{F} ,
 58 matrices of dimensions $n \cdot p$ and $p \cdot m$, respectively, where p is the number of factors, and \mathbf{E} is the residuals matrix
 59 of dimensions $n \cdot m$.

60

61 Unconstrained PMF, although it can lead to robust results, is usually insufficient when the sources are highly
 62 correlated or have very similar source profiles. In such cases, guiding the model by introducing a priori knowledge
 63 (common practice known as constraining the model) has been proven beneficial for the source deconvolution
 64 (Lingwall and Christensen et al. 2007, Belis et al. 2014, Dinh et al. 2025). However, it can still introduce

65 substantial bias in the solution (Via et al. 2022). Globally, the RMs cover the whole range of pollution sources
66 knowledge required prior to receptor modelling (Viana et al. 2008, Belis et al. 2013). A very strongly-constrained
67 RM is the Chemical Mass Balance model (CMB), which factorises the initial matrices with a totally fixed \mathbf{G} or \mathbf{F} .

68
69 Bayesian models represent a probabilistic alternative to the PMF framework. The first application of Bayesian
70 models in atmospheric source apportionment was introduced in Park et al. (2001, 2002) for Volatile Organic
71 Compounds (VOC) source apportionment. In this approach, the mass closure condition was taken to the Bayesian
72 framework and an autocorrelation prior, AR(1) (the first order autoregression formulation), was applied,
73 improving the solution given assuming independent \mathbf{G} components. The autocorrelation prior importance was
74 later reinforced in Rusanen et al. (2024) with a differently formulated autocorrelation prior. The latter shows the
75 added value of the Bayesian Autocorrelated Matrix Factorisation model (BAMF) compared with PMF across
76 different spectrometry-based PM synthetic datasets. The Bayesian Multivariate receptor modelling software
77 BNFA and bayesMRM (Park and Oh, 2021) were developed to provide user-friendly tools for Bayesian source
78 apportionment.

79
80 However, studies using the Bayesian Matrix Factorisation framework are still scarce. Some examples are Oh and
81 Park (2022), which employed a Bayesian RM to conduct multi-site source apportionment, and Zhang et al. (2023),
82 which performed NH_4^+ source apportionment through the Bayesian SIMMR package (Govan et al. 2023).
83 Bayesian models have also been used as a complement to standard RMs, as in Balachandran et al. (2013) where
84 a Bayesian model processing ensemble solutions of a chemical transport model and solutions of three RMs are
85 produced to then use it in CMB for production of final results. The Bayesian model focused on attributing the
86 proper weight to each of the ensemble components and improved the correlation of sources with their markers
87 compared to the traditional approach. Bayesian inference has also been used in Park et al. (2002) and Dai et al.
88 (2024) to generate spatially resolved source apportionment solutions adjusting the weights of each location
89 solution in a multi-site data scheme.

90
91 Thus, Bayesian Matrix Factorisation has become an effective and powerful tool for aerosol source apportionment.
92 However, to the authors knowledge, little attention has been given to improving the accuracy of chemical
93 composition profiles, i.e. \mathbf{F} components. This highlights the fundamental challenge in receptor modelling of
94 obtaining chemically distinct and interpretable source profiles from complex and mixed emission sources.
95 Moreover, it has been shown in Rusanen et al. (2024) that in BAMF, slight differences of \mathbf{F} can severely
96 compromise the quality of \mathbf{G} (Figure S2 in the mentioned article), hence, steps towards \mathbf{F} refining should result
97 in overall source apportionment method improvement. In this context, sparsity, defined as the property of a
98 dataset, model or solution in which only a limited number of elements are substantial contributions while most
99 are zero or close to zero, could be favourable for this problem. The accomplishment of sparse source fingerprints
100 could represent “cleaner” emission sources with less mixing among resolved factor profiles, since substituting
101 non-significant contributions in a factor by zeros might allow allocating more importance to the actually relevant
102 contributions of species in factors. This work aims to implement sparsity on chemical fingerprints in BAMF
103 aiming for a more accurate source apportionment. We introduce sparsity with the regularised horseshoe prior
104 (Piironen and Vehtari, 2017), which unlike other sparsity priors, enables regularisation of the sparsity strength,

105 and compare it with other sparsity priors, such as Lasso (Tibshirani et al. 2015) and Spike-and-slab (Andersen et
106 al. 2014). This model is then tested on three synthetic datasets with different complexity degrees and one real-
107 world dataset to depict the impact of sparsity and potential benefits of its implementation.

108 2. Methodology

109 2.1 Bayesian Matrix Factorisation

110

111 Bayesian Matrix factorisation models, like other RMs, are based on the chemical mass balance equation (Eq. 1).
112 Bayesian modeling approaches this problem probabilistically and bases the determination of the matrices, \mathbf{F} and
113 \mathbf{G} , the main parameters to determine, upon the assumptions imposed on the model, i.e. priors. Bayesian
114 factorisation forces the decomposition through modelling the \mathbf{X} matrix components as a Gaussian with center on
115 the “noise-free data matrix” \mathbf{Z} (matrix of dimensions $n \cdot m$) and a standard deviation given by the positively-defined
116 uncertainty matrix (Eq. 2). The matrix σ (positive matrix of dimensions $n \cdot m$) represents the uncertainties of the
117 measurements. The matrix \mathbf{Z} is, in turn, the product of the time series and profiles submatrices, \mathbf{G} and \mathbf{F} ,
118 respectively, and (a) is rewritten as:

$$119 \quad X \sim N(\mathbf{Z}, \sigma) = N(\mathbf{G} \cdot \mathbf{F}, \sigma) \quad (2)$$

120 where N represents the normal distribution. With that formulation, the measurements matrix \mathbf{X} is modeled into a
121 Gaussian distribution whose centre is the $\mathbf{G} \cdot \mathbf{F}$ product matrix and its standard deviation is the uncertainty matrix
122 σ . In turn, one introduces certain restrictions on the \mathbf{F} , \mathbf{G} matrices characteristics in the form of priors. Whilst \mathbf{G}
123 is not given any prior and is sampled then by default from a uniform distribution, \mathbf{F} is modelled as a Dirichlet
124 distribution to ensure positivity, with the sum of its components being equal to 1 (2):

$$125 \quad F_k \sim \text{Dirichlet}(1_m) \quad (3)$$

126 With these \mathbf{F} requirements, profiles represent the normalised contribution to the spectra of one source. Usual
127 notation for indices used hereinafter are i, j, k for elements in the range $(1, \dots, n)$, $(1, \dots, m)$, and $(1, \dots, p)$, for the
128 timestamps, species, and factors, respectively. It is worth noting that PMF applies the normalisation of profiles
129 after a \mathbf{F} , \mathbf{G} solution is found, not as a model prior as done in BAMF. The PMF generates mass-loaded \mathbf{F} , \mathbf{G}
130 solution matrices, which are reweighted to provide a normalised \mathbf{F} and a mass-loaded \mathbf{G} . In the Bayesian models
131 used in this study, the normalisation of \mathbf{F} is inherent to the model by design. The different formulations eventually
132 provide normalised \mathbf{F} and mass-weighted \mathbf{G} , with unlikely affectations due to the normalisation procedure. The
133 model configuration given by (2) and (3) will be referred to as Bayesian Factorisation model (BMF) and represents
134 the analog of PMF in the Bayesian framework. All models used in this manuscript are outlined in Table 1.

135

136 On top of this structure, Rusanen et al (2024) proposed an autocorrelation prior for \mathbf{G} which should account for
137 the inherent autocorrelation of air pollutant sources in time. The imposition of autocorrelation in \mathbf{G} entails that
138 two consecutive measurements should be more similar than two measurements apart in time, and that the similarity
139 should fade with the temporal gap between them. This property is particularly advantageous for atmospheric
140 pollution dynamics which, generally, are expected to exhibit temporal smoothness rather than abrupt fluctuations.
141 The formulation of the autocorrelation prior for \mathbf{G} is given by (4) and includes two more modelling parameters, α

142 (positive vector or dimension p) and β (positive vector or dimension p), which regulate the similarity of one \mathbf{G}
 143 component with the previous one as follows:

$$144 \quad G_{i+1,k} \sim C^+(G_{ik}, \alpha_k \cdot \Delta t_i + \beta_k) \quad (4)$$

145 where $i \in (2, \dots, n-1)$, C represents the Cauchy distribution and $+$ represents positive real numbers. This prior
 146 centers the $i+1$ th component distribution in the i th component with a distribution width that linearly depends on
 147 the temporal gap between these two timestamps. Hence, the more temporally-separated two consecutive points
 148 are, the less correlated they are expected to be. The Cauchy distribution was chosen due to its heavier tails which
 149 enable more probable jumps between consecutive i 's than a Gaussian distribution (Gelman et al., 2013). This
 150 flexibility could be convenient for real-world datasets which are affected by measurement gaps. The coefficients
 151 α and β are source-dependent to allow for source-dependent correlation degrees. The model which introduces this
 152 prior to BMF is called Bayesian Autocorrelation Matrix Factorisation model (BAMF, Rusanen et al. 2024).

153 **2.1.1 The horseshoe prior**

154 Here we propose a sparsity enforcement into the profiles matrix, intending to remove small contributions of
 155 irrelevant species for a given factor. The introduction of sparsity in BAMF involves the addition of several
 156 hyperpriors in the \mathbf{F} prior to implement the shrinkage mechanism. In this study, we used the regularised horseshoe
 157 prior (Piironen and Vehtari, 2017), which is a global-local complex of hyperpriors, i.e. the shrinkage power is
 158 both regulated globally source-wise and \mathbf{F} -component-wise. The idea behind this prior is that species with very
 159 small contributions to a factor are shrunk toward zero through an automatic shrinkage mechanism, whereas species
 160 with substantial support from the data are largely unaffected. The regularised horseshoe (HS) prior implemented
 161 in \mathbf{F} in the BAMF scheme as

$$162 \quad F_{kj} = \mu_{kj} \cdot \tilde{\lambda}_{kj} \cdot \tau_j, \quad (5)$$

163 where μ (matrix of dimensions $n \cdot m$) represents the \mathbf{F} matrix without the horseshoe prior. μ , in turn is defined as
 164 a standard Cauchy distribution prior as

$$165 \quad \mu_{kj} \sim C^+(0, 1), \quad (6)$$

166 where τ (vector of dimension p) represents the global shrinkage parameters and

$$167 \quad \tau \sim C^+(0, \tau_0 \cdot \sigma_{HS}), \quad (7)$$

168 where the parameter τ_0 can be regulated by the user to regulate the overall shrinkage power and σ_{HS} is sampled
 169 from an uniform distribution. The hyperparameter $\tilde{\lambda}_{kj}$ applies the local shrinkage to

170 as

$$171 \quad \tilde{\lambda}_{kj} = \sqrt{\frac{c^2 \cdot \lambda^2}{c^2 + \lambda^2 \cdot \tau^2}}, \quad (8)$$

172 where

$$173 \quad c^2 \sim \Gamma^{-1}(0.5 \cdot \text{slab_df}, 0.5 \cdot \text{slab_df}) \quad (9)$$

$$174 \quad \lambda \sim \text{slab_scale} \cdot C^+(0, 1) \quad (10)$$

175 both combined providing the characteristic shrinking horseshoe shape. Here, λ is a model parameter of dimensions
 176 $n \cdot m$ which after regularisation becomes is denoted as $\tilde{\lambda}$ Further description of the horseshoe implementation on
 177 BAMF can be found in Section S2.1, and the prior derivation and details in Piironen and Vehtari (2018). The
 178 distribution parameters τ_0 , σ_{HS} , and slab_df , slab_scale were tested and results did not show significant sensitivity
 179 to their variations, so we keep the default ones as provided in Piironen and Vehtari (2018) in their available shared

180 codes. The models with the horseshoe (HS) priors are hereinafter marked with "+HS". Figure S1 shows a
 181 schematic diagram of the matrix decomposition through BAMF.

182

183 In order to assess the amount of sparsity of a dataset or a solution, we used the Gini coefficient (Gini et al., 1936),
 184 which assesses the inequality over a distribution as follows:

$$185 \quad Gini = \frac{\sum_{i=1}^n (2 \cdot i - n - 1) \cdot x_i}{2 \cdot n \cdot \sum_{i=1}^n x_i} \quad (11)$$

186 where x values are sorted in ascending order, and n is the number of elements in x . It is a proxy for how deviant
 187 a dataset is from the total equality amongst its components. Since it quantifies the inequality, it can be a proxy for
 188 sparsity; if some values are high and the others are zero, $Gini \approx 1$ (great inequality), if all values are equal, $Gini$
 189 $= 0$. Also, the solution-to-truth Gini values ratio will be discussed throughout the analysis, referred to as “Gini
 190 ratio”. To evaluate if the sparsity is enforced precisely where it should, an additional metric has been applied
 191 called “zero truth sum”. This metric sums up the modelled contributions of the null species in the truth profiles.

192 **2.1.2 Alternative factorisation methodologies**

193 **BAMF-AR1.** There is an alternative formulation for the autocorrelation prior as introduced in Bayesian models
 194 by Park et al. (2001). The AR(1) autocorrelation prior is the first degree polynomial expansion of the
 195 autoregressive models and it proposes a linear progression of $G_{i+1,k}$ from $G_{i,k}$. We introduce AR(1) in the Bayesian
 196 framework as

$$197 \quad G_{i+1,k} \sim N(\alpha_k \cdot G_{i,k} + \beta_k, \gamma_k) \quad (12)$$

198 In this formulation, the $i+1$ -th point stems from a Gaussian distribution centered linear combination on the i -th
 199 point with source-dependent slopes (α) and intercept (β), and width (γ). Although, unlike (4), it disregards the
 200 decrease of correlation between gapped consecutive points, this prior allows for source-specific time series trends,
 201 which would be beneficial for certain source description. The model which introduces this prior to BMF will be
 202 called BAMF-AR1.

203

204 **BAMF-GS.** Another formulation is introduced, switching the Dirichlet distribution to the matrix \mathbf{G} instead of \mathbf{F}
 205 (6). This swap should allow \mathbf{F} to retain the \mathbf{X} matrix mass and could potentially help deconvoluting profiles due
 206 to the upweighting of the chemical profiles.

$$207 \quad G_k \sim Dirichlet(1_n) \quad (13)$$

208 Thus, the \mathbf{G} presents two priors, the dirichlet distribution and the autocorrelation prior, whilst \mathbf{F} is sampled from
 209 the default uniform distribution. This model will be called hereinafter BAMF-GS as short from BAMF-G simplex,
 210 since a simplex is the set of positive vectors that sum to one hence it is the natural geometric structure for the
 211 Dirichlet distribution to sit on. This model structure, nevertheless, does not allow for a horseshoe prior application,
 212 since due to the factors mass now incorporated in \mathbf{F} , the coefficients will be very distinct from zero and the
 213 horseshoe prior will not perceive them as potential signals to sparsify.

214

215 **CMB.** Lastly, a Bayesian formulation of the CMB model was employed in order to test the horseshoe prior
 216 capacities with the most proper factorisation possible. This model is mainly analogous to CMB in the Bayesian
 217 framework, but the \mathbf{G} matrix was fixed with the truth time series. Hence the model only had to determine the \mathbf{F}
 218 components distributions to match the factorisation condition (2) given the truth \mathbf{G} .

219 2.1.3 Solver and Hamiltonian-Monte Carlo Markov Chain

220 All Bayesian models were compiled and run in STAN (Carpenter et al. 2017), a probabilistic programming
221 language developed for Bayesian modelling. STAN solves Bayesian inference through the Hamiltonian Monte
222 Carlo (HMC) algorithm based on Markov Chain Monte Carlo methods (MCMC). HMC uses an approximate
223 Hamiltonian dynamics simulation with the Metropolis acceptance/rejection criterion and a no-U-turn sampler
224 (NUTS, Hoffman and Gelman, 2014). For the sake of brevity, we present only the essential concepts here,
225 directing readers to Carpenter et al. (2017), Gelman et al. (2014), STAN Manual (2025) and references therein for
226 comprehensive information.

227 The objective of the inference is to retrieve the parameters of the model, primarily \mathbf{F} and \mathbf{G} but also all the other
228 defined parameters (τ , λ , α , β). These are sampled from their posterior distributions, constructed from the priors
229 and the data introduced. In the Hamiltonian analogy, the evolution of these parameters across samples is computed
230 as the trajectory of a fictitious particle. This particle moves through the parameter space driven by random
231 momentum in all directions. This approach avoids the random-walk behavior of simpler sampling methods and
232 enables faster convergence. The trajectory is hence simulated using a discretized approximation, and candidate
233 positions are accepted or rejected according to the Metropolis criterion (Metropolis et al. 1953). Accepted
234 positions correspond to plausible parameter values (of the \mathbf{F} , \mathbf{G} , τ , λ , α , and β parameters in our case) values given
235 both the model assumptions and the data. This process provides a distribution over samples of possible solutions
236 from which confidence intervals of each of the model (hyper)parameters can be extracted. A set of samples is
237 called a chain, each of them initialised with a different seed to explore the solution space more broadly. In order
238 to initialise the model parameters more effectively, the maximum a posteriori (MAP) point parameters solution
239 estimated by STAN is used through the LBFGS algorithm (Liu and Nocedal, 1989). Even if this approach makes
240 the parameter sampling process much more efficient, solutions might have multiple local maxima, and MAP will
241 initialise the models based only on one of those. This highlights the importance of using different seeds to explore
242 the solution space more widely. Since the early iterations of each Markov chain are typically influenced by the
243 starting and may not represent samples from the true posterior distribution, we discarded the first half of the
244 samples from each chain. Different settings were used according to the type of experiment, shown in Table S1.
245 The number of chains is consistent with standard practice in Bayesian modeling, and the number of samples was
246 increased beyond commonly adopted values (e.g., 1000) in order to improve solution stability. As seen in Table
247 S1, the more complex the datasets are, the more time BAMF+HS takes to run. Since the BAMF+HS running times
248 are high at this development stage, BAMF+HS might currently be more adequate for exhaustive source
249 apportionment refinement than real-time monitoring.

250 In order to evaluate the convergence of a solution to the target posterior distribution, the potential scale reduction
251 factor (\hat{R} , Gelman and Rubin, 1992) is used. This coefficient compares the variance within chains and between
252 chains of the Z matrix, hence if chains converge, $\hat{R} \approx 1$, values of $\hat{R} \gg 1$ imply chain divergence and values of $\hat{R} \ll 1$
253 imply sampling divergence in chains. The convergence of all runs has been assessed using standard Bayesian
254 diagnostics, including visual inspection of trace plots, and the effective sample size and \hat{R} statistics, and all
255 experiments shown in the manuscript fall within satisfactory stability ranges for these criteria.

256

257 2.2 PMF

258 In PMF (Paatero and Tapper, 1994), equation (1) is solved through the ME-2 solver (ME2, Paatero, 1999) based
259 on the weighted means squares minimization of the quantity:

$$260 \quad Q = \sum_{j=1}^m \sum_{i=1}^n \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2 \quad (14)$$

261 PMF was implemented on all datasets unconstrainedly through the Source Finder software (SoFi version 9.5,
262 Canonaco et al. 2013) with 100 runs which are posteriorly sorted as in BAMF. The number of runs may seem
263 compromising the PMF quality in comparison to the 4000-12000 samples per chain used in Bayesian models.
264 However, this comparison is misleading, since the factorisation space is indeed better explored by PMF, with 100
265 different sampling seeds, while only 4 seeds (chains) were used in BAMF-like models as usual procedure in
266 Bayesian modelling for the sake of computational resources.

267

268 2.3 Pre- and post- processing for all models

269 Before model running, \mathbf{X} and $\boldsymbol{\sigma}$ are normalised to use consistent scales of all priors and posteriors for Bayesian
270 models. The normalisation is based on ensuring a mean of $\mathbf{X} = 1$.

$$271 \quad X^* = X / f_{norm} \quad , \quad \sigma^* = \sigma / f_{norm} \quad \text{where} \quad f_{norm} = \sum_{i,j} X_{ij} / (n \cdot m) \quad (15)$$

272 After the factorisation this normalisation is reverted converting the normalised matrices, hereinafter referred as $\bar{\mathbf{G}}$,
273 $\bar{\mathbf{F}}$, to the properly-scaled \mathbf{G} , \mathbf{F} matrices.

274

275 In the model outcomes, the factor ordering in the matrices is random in the model results, hence, the solution
276 factors must be sorted. Here, as in Rusanen et al. (2024), we used the Hungarian algorithm (Kuhn, 1955) to sort
277 the $\bar{\mathbf{Z}}_k$ components ($\bar{\mathbf{Z}}_k = \bar{\mathbf{G}}_{tk} \cdot \bar{\mathbf{F}}_{kj}$, i.e. each factor's normalised \mathbf{Z} submatrix). The metric to sort the components
278 is the Manhattan distance (i.e. the sum of the absolute differences of two Cartesian coordinates). All factors in
279 each chain of samples are then reordered upon the factor order of a small group of samples of that chain (the last
280 5, arbitrarily chosen) and, subsequently, one all-samples-averaged \mathbf{F}_k and \mathbf{G}_k are retrieved for each of the chains.
281 Then, the order of factors of each of the chains is sorted again in the same way in relation to the truth \mathbf{F} , to have
282 all sources equally sorted in all chains. Median and quantiles are computed over samples and chains to produce
283 the final solutions and uncertainties. This sorting process is also used for the PMF solution despite not being its
284 usual sorting approach for the sake of homogeneity in comparison to the Bayesian models.

285

286 The last step of the experimental process was to assess the model performance on the given dataset. The evaluation
287 of the performance should be based on: i. reconstruction performance, or the difference between \mathbf{X} and \mathbf{Z} ; ii.
288 similarity to truth, or environmental sensibility based on the apportionment of source tracers in case the truth is
289 not available; iii. computational performance. The reconstruction performance was assessed by checking the cell-
290 wise correlation between \mathbf{X} and \mathbf{Z} and checking the median and maximum of the absolute value of relative
291 deviations of \mathbf{Z} and \mathbf{X} with respect to the measurement uncertainty matrix $\boldsymbol{\sigma}$ ($|\mathbf{X}-\mathbf{Z}|/\boldsymbol{\sigma}$). The similarity to truth,
292 when available, is tackled by comparing the median ratio between modelled \mathbf{G} and truth (\mathbf{G}/\mathbf{G}_0), the Pearson
293 correlation for the \mathbf{G} matrix ($\mathbf{G} \rho$), and the Spearman correlation for \mathbf{F} amongst models ($\mathbf{F} \rho$). The Spearman
294 correlation coefficient for the factor profiles was chosen due to the expected non-linearity of the comparison and

295 likely presence of outliers. These comparisons, and especially when the ground truth is not available, need to be
296 accompanied by visual inspection of the solution quality, looking for resemblance with known environmental
297 sources. The models accounting for sparsity will be also compared upon the aforementioned Gini metric and,
298 when truth is available, the Gini ratio with truth and the “zero-truth sum”. Computational performance assessment
299 will be based on the metrics of convergence metrics of the Hamiltonian-Montecarlo Markov chain methods
300 embedded in STAN software (e.g. \hat{R}).
301

302 **2.4 Datasets**

303 The datasets created for model experimentation can be divided into synthetic and real-world datasets. Synthetic
304 datasets are artificially created with the purpose of knowing the \mathbf{F} , \mathbf{G} , to test model accuracy retrieving these
305 matrices with respect to the *truth* and these have been widely used for source apportionment validation in the last
306 decades (Park et al. 2002, Brinkman et al. 2006; Belis et al. 2015; Via et al. 2022; Rusanen et al. 2024). In order
307 to challenge the models gradually, we created three synthetic datasets with increasing degrees of complexities
308 (toy, offline, online ACSM synthetic datasets). Additionally, a real-world chemically sparse dataset was also used
309 to test the results. Although the truth factorisation is unknown and the results cannot be directly verified, the
310 model's factorisation can be assessed environmentally or based on indicators on the goodness of fit. The different
311 datasets have different levels of sparsity, as can be seen in Table 2, that the models with the horseshoe prior should
312 aim to replicate. The time resolution of modelled OA sources, used both in the chemically-sparse toy dataset and
313 the chemically less sparse datasets, is 1 hour. The time resolution of offline datasets, used in the chemically sparse
314 synthetic offline dataset and the chemically-sparse real-world offline dataset is 1 day.

315 **2.4.1 Chemically-sparse toy dataset**

316 A simplistic synthetic toy dataset was designed as a deliberately simplified test case to perform basic control and
317 performance tests, rather than to reproduce any realistic atmospheric scenario. It was devised by creating three
318 very simple and sparse profiles and using three time series (HOA, SOA_{bio}, BBOA) from modelled source time
319 series of the city of Zurich (Rusanen et al. 2024, time resolution of 1h) in order to test how sparsity priors act on
320 very uneven species contribution. Although it is based on ACSM-like time series and therefore reflects some of
321 the temporal properties of such measurements, the three included sources do not represent combinations that
322 would be expected in a real-world environment since this toy dataset is intended solely for methodological testing
323 purposes. In addition, the source profiles were intentionally designed to be highly simplified in order to facilitate
324 an immediate visual assessment of the model fitting. For these reasons, the extracted components were not
325 assigned environmental labels, but were instead referred to generically as Factor 1, Factor 2, and Factor 3.
326

327 Then, \mathbf{F} and \mathbf{G} were multiplied to generate \mathbf{Z} , and some gaussian error with standard deviation σ was added to
328 each component to generate a realistic \mathbf{X} matrix. The uncertainties matrix σ was designed as a sixth of the \mathbf{X} values
329 plus Gaussian noise. With this arrangement, the models can be applied conventionally to the \mathbf{X} , σ matrices and
330 the modelled \mathbf{F} and \mathbf{G} , can be compared to the original truth \mathbf{F} , \mathbf{G} , which will be referred hereinafter as \mathbf{F}_0 and \mathbf{G}_0 ,
331 displayed in Figure S2.

332 2.4.2 Chemically-sparse synthetic offline dataset

333 We created a synthetic offline filter dataset, mimicking the filter-based measurements input matrices, in order to
334 test the accuracy of the models in these kinds of datasets. This dataset mimics the concentrations on the coarse
335 fraction ($PM_{10} - PM_{2.5}$) as collected by a high-volume sampler on the Zurich-Kaserne site (Grange et al. 2021)
336 including the following chemical species: OC, Al, Na, Mg, Cl, K, Ca, S, Fe, Cu, Zn, Mn, Sb, Ba, mannitol,
337 arabitol. In the original real-world dataset, data obtained with two series of samples (PM_{10} and $PM_{2.5}$) were
338 subtracted in order to focus on the coarse source apportionment, since the main emission sources of these elements
339 and organic species stem from mechanical processes leading to major coarse models. It was created by crafting
340 first the \mathbf{F} and \mathbf{G} , then multiply them and creating \mathbf{X} and σ . The \mathbf{F} matrix was slightly modified from that proposed
341 in Manousakas et al. (2025), making the chemical profiles slightly sparse by zeroing the non-relevant species in
342 each of the factors (dust, traffic, salt, coarse biological). The \mathbf{G} matrix was composed of the time series of:

- 343 - Dust: modelled PM_{10} dust (Vasilakos et al. in prep.) converted to coarse with the $Al_{PM_{10}}$ vs. PM_{10} ratio
344 from Grange et al. (2021).
- 345 - Traffic: modelled PM_{10} copper (Upadhyay et al. 2025) converted to coarse with the $Cu_{PM_{10}}$ vs. PM_{10}
346 ratio from Grange et al. (2021).
- 347 - Salt: coarse Na+Cl (Grange et al. 2021) converted to PM concentrations and multiplied by an arbitrary
348 number (3 in this case match the concentrations of the sea salt factor in the original dataset).
- 349 - Coarse biological: coarse Arabitol+Mannitol (Grange et al. 2021) converted to PM concentrations and
350 multiplied by 3, similarly as for the salt factor.

351 This dataset will be called “offline synthetic dataset”. Another more simplistic dataset was prepared similarly but
352 using Al and Cu for dust and traffic factors, respectively, in the same way as in the salt or coarse biological factors,
353 i.e. omitting the use of modelled data. This dataset will be hereinafter named “Purely-measurement-based offline
354 synthetic dataset” and its modelling results will be described in section 3.2. Once the \mathbf{F} and \mathbf{G} matrices were
355 created, \mathbf{X} was calculated by their multiplication and the addition of Gaussian noise with amplitude σ . The
356 uncertainties matrix σ was generated as in Grange et al. (2021) multiplied by 2 to balance the signal-to-noise ratio
357 to the datasets in Manousakas et al. (2025). The matrices \mathbf{F} , \mathbf{G} of this dataset are displayed in Figure S3.

358

359 2.4.3 Chemically sparse real-world offline dataset

360 A real-world dataset was employed to test the current models applicability in campaign measurements. This
361 dataset was originally used for source apportionment in Manousakas et al. (2025) and Grange et al. (2021) and
362 consists of $PM_{10-2.5}$ samples at five Swiss National Air Pollution Monitoring Network (NABEL): Basel, Bern,
363 Magadino, Payerne, and Zurich. The measurements were taken in the June 2018 - July 2019 period every fourth
364 day and using Digital high-volume samplers. During the sampling campaign PM_{10} and $PM_{2.5}$ were collected and
365 the respective concentrations were subtracted to generate the coarse ($PM_{10-2.5}$) concentrations. These samples
366 include: i. OC concentrations, measured through the thermal optical transmission (TOT) EN16909 method with
367 the EUSAAR2 temperature protocol; ii. elemental concentrations (Al, Fe, Cu, Zn, Mn, Sb, Ba, Sr, Bi, Pb)
368 measured by inductively coupled plasma atomic emission spectrometry (ICP-AES) and inductively coupled
369 plasma mass spectroscopy (ICP-MS); iii. water soluble inorganic ion concentrations (Ca^+ , Cl^+ , Mg^+ , K^+ , Na^+),
370 determined by ion chromatography (IC); iv. Organic species (mannitol, arabitol) determined by a high-

371 performance liquid chromatographic method followed by pulsed amperometric detection (HPLC-PAD). The
372 uncertainties of these species were calculated as in Grange et al. (2021).

373 2.4.4 Chemically less sparse synthetic online ACSM datasets

374 With the aim of recreating more complex real-world datasets to test the models, we generated 6 datasets for four
375 European cities: Krakow, Milan, Paris, and Zurich. The objective was to recreate OA matrices as given by a mass
376 spectrometer instrument like Q-ACSM, for which there are plenty of real-world source apportionment studies in
377 the literature. The **G** matrix was created from OA sources time series generated through the regional air quality
378 model CAMx (Comprehensive Air Quality Model with Extensions) as previously published by Jiang et al. (2019).
379 The five sources of these datasets were hydrocarbon-like OA (HOA), related to traffic emissions, biomass burning
380 OA (BBOA), biogenic SOA (SOA_{bio}), biomass burning SOA (SOA_{bb}), and traffic SOA (SOA_{tr}). To ensure
381 seasonal representativity while keeping computational costs low, datasets included the first two weeks of every
382 second month of 2011 (January, March, ...). The relative concentrations of these datasets are shown in Figure S5.
383 This figure shows the highest seasonal OA variation for the city of Milan and the lowest for Zurich. In terms of
384 sources, the most seasonally stable sources, overall, are HOA and SOA_{tr} in contrast to the remarkable variability
385 of BBOA and SOA_{bio}. The profiles used to create the species matrix **F** were those in Table S2 for primary sources
386 (HOA, BBOA). For secondary sources, the profiles from the European megacity dataset presented in Rusanen et
387 al. (2024) were used for the Zurich city, which were slightly perturbed for the other cities due to the limited
388 availability of these sources' profiles in the literature.

389
390 The **X** matrix was obtained by multiplying the **F** and **G** submatrices and adding Gaussian noise. The procedure to
391 calculate the error matrix for such datasets is described in Via et al. (2022) and the dataset used to calculate the
392 error matrix is that from the Zurich site, which ranges from February 2011 until December 2011.

393
394 Lastly, a sensitivity analysis was carried out by slightly modifying the original **F**, **G** matrices upon which the **X**,
395 **σ** matrices were subsequently created. The first Zurich dataset (period 01/09/2011 - 14/09/2011) was used for this
396 purpose and we chose to perturbate one factor only (HOA). The **F**, **G** submatrices were perturbed independently
397 upon the expression:

$$398 \quad G_{HOA}' = G_{HOA} \cdot N(1, \sigma') \quad F_{HOA}' = F_{HOA} \cdot N(1, \sigma') \quad (14)$$

399 where we used $\sigma' = [0, 0.1, 0.2, 0.3, 0.4, 0.5]$ to create different degrees of perturbation. The profiles in **F** were
400 normalised after that process. It must be noted that the perturbation is more relevant on **F** than in **G** since a given
401 σ' in the aforementioned range is more comparable and impactful on the profile contributions, bounded to 1, than
402 on the unbounded time series timepoints. Consequently, within this framework, we obtained 6 **G**-perturbed and 6
403 **F**-perturbed input matrices. Both BAMF and BAMF+HS models were run with all these input matrices and their
404 subsequent HOA results were compared to the original truth in order to comprehend the sensitivity of the models
405 upon time series and profile perturbations.

406 3. Results

407 3.1 Chemically sparse synthetic toy dataset

408 Here, we introduce the evaluated models relying on unrealistically simplified toy data with the purpose of
409 showcasing the performance of the horseshoe prior introduction to BAMF (Figure S2) and the alternative
410 factorisation methodologies, which are discussed in SI Section C.1.

411

412 In the first evaluation step, we assess the performance of the horseshoe prior under the assumption that the source
413 matrix \mathbf{G} is known, in order to isolate its effect on the estimation of \mathbf{F} . Figure 1 shows the distribution of each \mathbf{F}
414 component for CMB with and without the horseshoe prior (CMB, CMB+HS, respectively, Table 1). The
415 distributions shown account for all the variability across samples of each \mathbf{F} component for both models, and the
416 truth is shown as a marker in the x-axis since it is a point value to be compared to the centers of the distributions.
417 The presentation of the CMB and CMB+HS distributions aims to demonstrate the sparsity-inducing role of the
418 horseshoe prior, which enforces shrinkage of the \mathbf{F} component toward zero; this effect is more readily discernible
419 when a strongly guided \mathbf{G} matrix is used to isolate the evidence of sparsity. Figure 1 showcases the horseshoe
420 prior power to generate sparsity in \mathbf{F} components, shrinking more strongly the lowest signals to zero than CMB
421 and, as a consequence, enlarging the most prominent signals. Table 3 shows how the Gini metric is consistently
422 higher for CMB+HS with respect to CMB, supported by a higher Gini ratio and lower zero truth metric reflecting
423 the sparsification of profiles and higher similarity to truth. The RMSE compared to the truth for the profiles
424 improved with the horseshoe prior applied for all three factors (for CMB and CMB+HS, respectively: $1.2\text{e-}04$,
425 $3.8\text{e-}05$ for F1; $1.86\text{e-}04$, $5\text{e-}05$ for F2; $3.3\text{e-}05$, $1\text{e-}05$ for F3). Hence, the sparsity introduced in \mathbf{F} through the
426 regularised horseshoe prior successfully improved the profile description of the solution.

427

428 In the next evaluation step, we test the various models assuming no prior knowledge. Figure 2 shows the results
429 of PMF, BAMF, and BAMF+HS models on the toy dataset and Table 3 shows their factorisation performance
430 and comparison to truth metrics. In terms of factorisation, median relative errors are better for BAMF+HS and
431 BAMF than for PMF, but their maximum errors are higher and the Pearson coefficients slightly lower, all this
432 entailing comparable factorisation performances. All models generally adapt well to the truth features, but they
433 present non-negligible differences. PMF results better resemble the truth in terms of \mathbf{G} R^2 , but it is the model
434 whose G/G_0 differs from 1 the most, accumulating the greatest error (2.64), followed by BAMF (2.10), while
435 BAMF+HS exhibits the smallest deviation (0.81), indicating the highest overall accuracy. In terms of profiles, the
436 BAMF+HS model is the closest to the truth both in terms of ρ and R^2 , especially for the second and third factors
437 for which the sparsity introduction results are advantageous with respect to BAMF results. Consistently, the Gini
438 ratios of the inferred solutions relative to the truth are markedly closer to unity for BAMF+HS (range 0.40–0.93)
439 than for PMF (0.45–0.64). The sparsity effects can also be seen in Figure S5, in which the horseshoe shrinkage is
440 evident for the low m/z s allowing in turn the larger m/z s to retain more mass, hence resembling better the truth
441 profiles. Taken together, these results indicate that BAMF+HS not only promotes sparsity, but does so in a
442 chemically consistent manner, leading to a more accurate mass apportionment across factors, despite a slightly
443 reduced time-series correlation for the third factor. However, the BAMF+HS could not shrink down the lowest
444 signals in Factor 1, likely because their contribution estimated by the mass balance and the autocorrelation
445 restrictions of this model made it unclear for the horseshoe to shrink them down completely. With this result, this

446 toy dataset depicts the capacities and limitations of the horseshoe implementation on BAMF: it is capable to
447 sparsify effectively only the signals which are close enough to zero as given by the restrictions of the BAMF
448 model.

449 While other sparsity priors exist (e.g. Lasso and Spike-and-slab priors (Figure S6, Table 3, Table S3)), our tests
450 show that the BAMF+HS model is most effective in shrinking unnecessary contributors to \mathbf{F} . Hence this prior will
451 be used onwards. This is evidently portrayed by the Gini ratio, for which neither Lasso nor Spike-Slab achieve
452 the signal shrinkage that the BAMF+HS does. Also, neither BAMF+Lasso nor BAMF+Spike-and-slab managed
453 to sparsify the first factor. Additionally, different autocorrelation formulations were implemented with and
454 without the horseshoe prior, showing worse performance than BAMF or BAMF+HS, respectively, as discussed
455 in section SI C.1. This supports using the BAMF autocorrelation prior instead of the alternative AR(1) prior, \mathbf{G}
456 simplex formulation or lack of autocorrelation prior models, although these models are also tried on the other
457 datasets to further highlight this.

458 **3.2 Chemically sparse synthetic offline dataset**

459 This synthetic offline dataset was used to assess the performance of different models on a proxy representation of
460 atmospheric aerosol data, while maintaining the verifiability property inherent to synthetic datasets as described
461 in Section 2.4.2. We performed source apportionment of the \mathbf{X} matrix through the aforementioned Bayesian
462 models and PMF, obtaining 4 factors fingerprints and time series. The dataset used in this source apportionment
463 is expected to be much more sparse than ACSM-like datasets, hence it could better expose the capabilities and
464 added value of the sparsity prior.

465
466 To avoid initialisation failure, BAMF was run by initialising \mathbf{F} as a normal distribution to ensure a more sturdy
467 sampling. Model initialisation fails when no set of initial parameter values satisfying the model result in valid
468 Bayesian solutions, and are usually solved by imposing more informative priors constraints on the model
469 parameters. A t-test was run comparing the \mathbf{F} , \mathbf{G} factors from this slightly modified model and BAMF to ensure
470 their similarity. Its results passed the t-test for all factors except for one factor, although it presented a $R^2=0.9990$
471 correlation and only a 20% of quantitative difference with that BAMF factor. Hence, one can assume that the
472 model provides an acceptable level of agreement with BAMF, capturing the essential structure of the factors with
473 only very minor deviations.

474
475 Figure 3 presents the (a) time series (b) auto-correlation (c) profiles of the source apportionment solution for PMF,
476 BAMF, BAMF+HS, (d) additional comparison to truth metrics, and Figure 4 shows the histograms of the models
477 \mathbf{F} components estimation. The time series and autocorrelation show only slight differences between the models,
478 the PMF being the most different to the truth in all factors except the salt one, as supported in Figure 3 (d).
479 Amongst factors, the coarse biological source is the most poorly reconstructed. If accounting for the sum of all
480 factors $\mathbf{G} R^2$ s and \mathbf{G}/\mathbf{G}_0 , in the last row of Figure 3(d), the most accurate model is the BAMF+HS, followed by
481 PMF and then BAMF. In terms of profiles, the best overall model performance depends on the metric, \mathbf{F} Spearman
482 correlation coefficient being highest for BAMF and R^2 and cosine similarity correlation coefficients for
483 BAMF+HS. This fact, accompanied by Gini being the highest for BAMF+HS and the closest to 1 Gini ratio,
484 indicates that the extreme values of the profile (i.e. maximum and zeros species contributions) are closer to truth

485 for BAMF+HS, whose extreme contributions would be less relevant in the Spearman correlation coefficient.
486 Considering the Truth \mathbf{F} zeros sum metric, the horseshoe shrinkage is visibly sparsifying most of the low signals
487 whilst BAMF and PMF present non-zero contributions for species whose contribution in this factor is null. Hence,
488 the BAMF+HS model would effectively promote the profiles sparsity which it was intended for.

489

490 However, the favourable results of BAMF+HS in comparison to the other models could be a dataset-dependent
491 finding, related to the properties of the created synthetic dataset. The purely-measurement-based offline synthetic
492 dataset, whose performance statistics are shown in Table S4, shows that PMF overperforms BAMF+HS,
493 presenting slightly higher \mathbf{F} and \mathbf{G} R^2 and better G/G_0 . This could indicate that the optimal model selection might
494 be dataset dependent. However, the source time series of this very simplistic dataset are fully correlated with some
495 species time series, since they are used to generate factor time series, which makes it a very redundant dataset. In
496 this scenario, the source apportionment comparison might still be valid, but it is not the perfect showcase for RMs
497 testing due to the excessive source correlation with species. We found it valuable to present different model
498 performances on different datasets, which in atmospheric measurements, can suffer from artefacts complicating
499 the behaviour of some models.

500

501 In the same way, the alternative autocorrelation priors models were also tried and will be thoroughly discussed in
502 Section SI C.II. However, overall, the BAMF+HS model is the one providing the best source apportionment
503 results for this offline dataset, taking advantage of the sparsity to upgrade both profiles and time series accuracy.

504

505 **3.3 Real-world offline dataset**

506 To test the models on real-world data and identify their limitations for more complex datasets, we tested the
507 models in the real-world offline PM_{10} - $\text{PM}_{2.5}$ dataset described in Section 2.4.3. Since the truth is not accessible,
508 the model performance can only be assessed upon environmental, factorisation-related, and computational criteria.
509 For this dataset, BAMF and BAMF-AR1 models presented initialisation issues preventing them from properly
510 launching the models. To avoid this issue and make the model more robust, we implemented a prior in \mathbf{F} so that
511 its components are drawn from Gaussian distributions centered at zero and with a standard deviation of 1 so that
512 we restrict values to be bounded to 1. This modification was not needed for the other models, which did not present
513 initialisation issues.

514

515 Source apportionment results for PMF, BAMF, and BAMF+HS are shown in Figure 5 and Table 4. Figure 6
516 shows the \mathbf{F} distributions for these models, as a detail of Figure S9 (a). Figure S6, S9 (a) display very similar
517 results for PMF, BAMF, BAMF+HS both in terms of \mathbf{F} , \mathbf{G} , and reconstruction metrics, and only some differences
518 can be perceived for PMF, while BAMF and BAMF+HS histograms are almost overlapping in Figure 6. However,
519 the BAMF+HS profiles present a remarkable difference in terms of sparsity as seen in the \mathbf{F} Gini metric, which is
520 mostly the highest for BAMF+HS or equal, except for the biological factor for which PMF is slightly higher. For
521 some species, the relative \mathbf{F} components apportionment is more strongly suppressed by BAMF+HS than by
522 BAMF or PMF, hence, their contribution on other profiles can be larger. This is clearly visible, for instance, for
523 OC, Mg^+ , K^+ , S^+ , or mannitol, which are zeroed in the Salt factor and consequently are larger on the factors where

524 these species are relevant. This is more evidently depicted in Figure S9 (a) and Figure 6, where the distribution of
525 **F** components is shown. For the aforementioned species, the horseshoe effect can be seen in the distribution,
526 whilst BAMF and PMF are further from zero. This result thus highlights the potential benefits of sparsity
527 introduction in matrix factorisation.

528

529 The application of other autocorrelation priors was not advantageous with respect to the regular BAMF
530 autocorrelation and even worsened the shrinkage power of the horseshoe prior as discussed in SI C.III.

531 **3.4 Chemically less sparse synthetic online ACSM datasets**

532 The next step was to test these models on more realistic synthetic datasets. For that purpose, 6 datasets for 4
533 European cities (a total of 24 datasets) were designed with 5 factors in each of them (section 2.4). We applied the
534 8 models under discussion (PMF, BMF, BMF+HS, BAMF, BAMF+HS, BAMF-AR1, BAMF-GS) to the 24
535 synthetic datasets and computed the summary statistics (the median of the ratios of **G** over the truth **G**, \mathbf{G}/\mathbf{G}_0 , the
536 Pearson correlation of **G** with truth, $\mathbf{G} \rho$, and the Spearman correlation of **F** with truth, $\mathbf{F} \rho$). All metrics over cities,
537 datasets and sources are presented in Table S5, and an example for one site (Zurich) and one dataset (dataset 0,
538 from 01/01/2019 to 14/01/2019) is shown in Figure S11 as an example of the results obtained by the three models
539 in 1 out of the 24 datasets.

540

541 Figure 7 shows the model summary statistics over the 6 generated datasets for the four cities and Figure S12 shows
542 the factor-dependent statistics. In this case, the (not-squared) Pearson correlation coefficient was used to compare
543 the results of the ACSM-like datasets more easily to those presented in Rusanen et al. (2024), which used this
544 metric. Figure 7 shows a good agreement between models and the truth, with most solutions with correlations
545 with truth for **F** and **G** above 0.7, similarly to Rusanen et al. (2024). However, there are clear differences amongst
546 models and cities. PMF is performing worse in comparison to the Bayesian models, including BMF, the Bayesian
547 analog to PMF in all datasets except for Milan. As shown in Table S5, PMF presents the highest $|\mathbf{Z}-\mathbf{X}|/\boldsymbol{\sigma}$, the
548 highest overestimations of **G**, and correlations of **G** and **F** are the lowest in comparison to other models except for
549 the Milan dataset. In terms of \mathbf{G}/\mathbf{G}_0 , the model providing the best results are BAMF, BAMF-AR1, BAMF-GS,
550 followed by their horseshoe versions. The BAMF+HS, presents slightly lower $\mathbf{F} \rho$, $\mathbf{F} R^2$, and the sparsity Gini
551 metric ratio is not close to one, entailing the horseshoe prior did not successfully implement sparsity and the **F**
552 accuracy did not improve. In terms of correlations with **F** and **G**, the models including the horseshoe prior present
553 higher dispersion within a city with respect to the models without sparsity terms. Considering all the parameters,
554 the models with the best overall performance are BAMF, BAMF-GS, and BAMF+HS.

555

556 Figure S13 shows the autocorrelation for lags 0-168 h (half of the monthly measurement period) for all the sources
557 and sites, displaying the cyclicity of the selected sources. In all cases, the short-term lags present very high
558 autocorrelation, entailing that the similarity on adjacent timestamps is very high and decays over longer periods.
559 Typically, and as presented on the figure, the autocorrelation of primary sources, with more marked daily cycles,
560 decays faster than secondary sources, which evolve more steadily due to their slower reaction to emissions. Whilst
561 HOA and BBOA present a very steady intradaily structure, with one or two maxima per day, the biogenic SOA
562 presents one peak per day and the other two secondary sources may or may not present marked daily cycles. This

563 different intra- and inter-daily structure amongst sources certainly challenges the models to resolve the source-
564 dependent characteristic.

565

566 Figure 8 shows the autocorrelation from truth and the model outputs correlate (Pearson coefficient of
567 determination) for each model and source in the 4 cities. Each dot represents one of the 6 datasets for each site,
568 and colors represent the different sources. The results show that all models present very high Pearson coefficient
569 ranges for **G** autocorrelations in comparison to truth except for PMF, which struggles with this dataset aspect due
570 to the lack of accounting of self-correlation. In general terms, the best captured correlation by all models is that
571 of SOA_{Bio}, with the most regular cyclical patterns. The SOA_{BB} and SOA_{Tr} autocorrelations seem to challenge the
572 models further due to more irregular patterns, and for some datasets, their autocorrelation is poorly modeled. POA
573 sources are generally accurately modelled, with HOA patterns slightly better captured than those from BBOA.
574 Regarding models, the ones with better performance are BAMF-GS, BAMF, and BAMF+HS, with only slight
575 differences between the last two. This observation suggests that the horseshoe prior addition does not significantly
576 reduce the autocorrelation power of the BAMF.

577

578 Regarding sparsity, Figure S14 depicts the lack of sparsity both for input and modelled data. This figure shows
579 the truth's 5 lowest m/z components as well as BAMF, BAMF+HS outcomes. The reference (truth) profiles do
580 not present zeros but very small signals, as do many ACSM-like profiles in the AMS spectral database (Ulbrich
581 et al. 2009). Both BAMF, BAMF+HS reflect this lack of sparsity, however, it could be expected that BAMF+HS
582 would decrease the contributions of the lowest components. However, the sparsity introduction was not achieved
583 as seen before in the lack of improvement of the Gini ratios. This lack of sparsity despite the enforcement through
584 the horseshoe prior can be explained by the complexity of the data, which due to chain divergence, hinders the
585 models performance. Figure S15 shows the model \hat{R} , a typical Bayesian metric to evaluate the precision of
586 Hamiltonian chains, computing the ratio between inter- and intra-chain variabilities. In any case results are very
587 close to the ideal value, 1, so the validity of all models' solutions is assured. However, this plot reflects the
588 deprecation of the solution with models when the horseshoe prior is applied. The horseshoe prior adds more
589 complexity to the **F** with three more parameters compared to non-sparsity models which could be the cause of the
590 increased model instability across chains.

591

592 Finally, a sensitivity analysis was run for the first Zurich dataset perturbing independently the original **F**, **G**
593 matrices to different degrees, monitoring the correlation of the modelled **F**, **G** matrices to the original truth (Figure
594 9). Subfigures (a) and (b) show how both in the case of the original **F** and **G** perturbations, the **F** accuracy drops
595 immediately and analogously for both models, with a more sudden decay for **G** perturbations. Contrarily, the
596 affectations in **G** (subplots (c) and (d)) are different for both models, with a steady decay for BAMF with **G**
597 perturbations and a non-clear trend for **F** perturbations, whilst BAMF+HS correlation rests insensitive to **F**, **G**
598 perturbations with an increasing/decreasing erratic behaviour. This result shows the reduced precision in **G** of
599 BAMF+HS in comparison to BAMF due to the chain divergence issue, which, in any case, does not severely
600 compromise its accuracy. This finding also explains the bigger variations for BAMF+HS with respect to BAMF
601 in all the metrics shown in Figure 8. Additionally, it showcases the general strong sensitivity of **F** determination
602 opposite to the general robustness of **G** upon general **X** matrix perturbatio.

4. Discussion

This study aims to explore further BAMF capabilities and the benefits introduced through additional priors and/or modifications of the current model structure as given by Rusanen et al. (2024). The introduction of sparsity in source apportionment models was of particular interest to provide more distinct and concise source profiles which can, in turn, improve the time series accuracy. However, in real-world applications, it may also remove small but relevant signals along with noise. Therefore, comparison with BAMF results is recommended, leaving it to the user to decide whether the method's use is appropriate for their case.

Firstly, the use of the simplistic toy dataset highlighted the added value of the sparsity introduction through the horseshoe prior in the totally constrained experiment. In this controlled setting, the ground truth structure is well defined, allowing the effect of sparsity to be clearly isolated and the method performance validated. However, for an unconstrained experiment, sparsity was proven remarkably advantageous, but subject to the underlying matrix factorisation results. That is, the horseshoe prior in BAMF+HS effectively suppresses weak signals of \mathbf{F} contributions as determined by BAMF, yet it fails to guide the model toward a more accurate or sparser solution when the initial BAMF estimate is suboptimal. Other sparsity priors, like Lasso and Spike-and-slab, were tried out but did not improve the regularised horseshoe performance.

The introduction of the regularised horseshoe prior in BAMF improved apportionment of offline synthetic and real-world datasets with respect to BAMF, promoting sparser profiles. The synthetic dataset comparison to truth was maximal for BAMF+HS, with sparser profiles and consequently better \mathbf{G} accuracy. Its application also proved advantageous for the real-world dataset, despite not being able to be compared to the truth. In this case, improvements are assessed through increased profile distinctness and internal consistency rather than absolute accuracy. The results show a sparsity effect which provides more distinct profiles in comparison to PMF and BAMF. This result encourages the usage of the horseshoe prior for sparsity introduction in datasets whose solutions are expected to be strongly sparse, such as elemental datasets.

Subsequently, in the more complex and realistic European datasets, the sparsity introduction could not be effectively enforced. Although solution quality was not substantially compromised, the profiles remained non-sparse after applying the prior. This is likely due to model instability arising from the higher complexity of these datasets, which is further aggravated by the addition of the horseshoe prior, as it requires sampling a larger number of parameters. Moreover, the inherent nature of ACSM datasets—characterized by highly correlated species—might also contribute to this limitation, since the model struggles to disentangle overlapping sources when variables are strongly interdependent. The higher chain divergence found for the *horseshoed* models causes a drop in solution precision due to different landings on the solution space depending on the chain. This issue could be reduced by selecting chains a-posteriori upon user-defined criterion as is practiced in PMF. This is further confirmed by the insensitivity to \mathbf{G} or \mathbf{F} perturbations that are visible for BAMF+HS but not for BAMF. Nonetheless, given that ACSM-like factor profiles exhibit low sparsity in the literature, the use of sparsity priors in these datasets is less justified. Also, because usually ACSM profiles obtained in chamber or ambient experiments are not usually sparse, as seen in Ulbrich et al. (2009), the BAMF+HS is not as pertinent in these kinds of datasets as for filter-based datasets.

643

644 The sparsity conceptual framework could also be brought into PMF through the pulling equations, which can
645 shrink down manually the expectedly low signals in a factor. However, this methodology requires that the user
646 indicates the species that are intended to be zeroed, which introduces user-subjectivity to the problem. The
647 BAMF+HS method, contrarily, acts globally, shrinking those species with lowest signals in favour of the matrix
648 factorisation, hence no user intervention is needed. This makes the approach more objective but also less targeted,
649 returning the factorization optimisation agency to the model. However, if the purpose were to enforce a shrinkage
650 of a certain species as in the PMF case, this feature could also be implemented through the horseshoe method with
651 minimal code modification.

652

653 The results of the other models tested (BAMF-AR1, BAMF-GS) did not show a significant improvement with
654 respect to BAMF. The BAMF-AR1 contains another autocorrelation to parametrisation (STAN Team, 2025)
655 which should allow for trend consideration, although this matter was not tackled in the current work and remains
656 to be validated in future studies. The BAMF-GS seemed to capture slightly better the \mathbf{G} variability in comparison
657 to BAMF in the online datasets, but led to worse correlation to truth in the offline synthetic dataset. Nonetheless,
658 it does not support enforcing sparsity in \mathbf{F} , thereby reducing its effectiveness for profile adjustments.

659 **5. Conclusions**

660 This study presents a sparsity introduction technique for the Bayesian Autocorrelated Matrix Factorisation model
661 (BAMF) which intends to condense source apportionment profiles removing noisy signals. The regularised
662 horseshoe prior, a tool to promote sparsity in datasets, is introduced in BAMF (BAMF+HS) in order to narrow
663 down the lowest signals in factor profiles while keeping the most significant ones regularised. The BAMF+HS
664 model is built in STAN, an open-source framework for statistical modelling with Hamiltonian-Montecarlo
665 Markov Chain sampling. In order to test the capabilities of the developed model, we generated three kinds of
666 synthetic datasets to compare the model factorisation outputs to the truth factors, namely Toy, offline, and online
667 synthetic datasets, each representing a progressively increasing level of complexity. Likewise, to confirm its
668 usability to real-world data, BAMF+HS was also applied to a multi-site filter dataset. Given the opportunity to
669 explore source apportionment with different types of datasets, we also tested other receptor models such as
670 Positive Matrix Factorisation (PMF) and other BAMF-like Bayesian models. In the Bayesian framework, we
671 tested a different formulation of the autocorrelation term (BAMF-AR1) and a permutation on the factorisation
672 matrix logic (BAMF-GS).

673 The main result highlights can be summarised as:

- 674 - BAMF+HS has been shown to be advantageous to introduce sparsity in factor profiles for offline datasets
675 and to not deprecate the solution for the more complex datasets mimicking Aerosol Chemical Speciation
676 Monitor (ACSM) data. Other sparsifying priors tried out were not as effective in low-signal shrinkage.
- 677 - The BAMF+HS performance towards truth profile reconstruction was higher than for BAMF and PMF
678 in the toy and offline synthetic datasets. Improving \mathbf{F} typically led to a more accurate determination of
679 \mathbf{G} , highlighting the strong interdependence between the two factorisation matrices.
- 680 - The real-world dataset also shows a better description of sources through BAMF+HS in terms of matrix
681 factorisation metrics and profile sparsification achievement.

- 682 - As shown in the toy dataset, the introduction of sparsity did not solve factorisation issues inherent to the
683 underlying factorisation model.
- 684 - The BAMF+HS model does not create sparsity in ACSM-like datasets, which are originally, indeed, non-
685 sparse. BAMF+HS is more unstable than BAMF for these more complex datasets as a result the higher
686 chain divergence during Hamiltonian-Montecarlo Markov Chain sampling as suggested by the \hat{R} metric.
687 However, the effects of the horseshoe prior do not affect the overall performance of BAMF or its
688 autocorrelation accuracy.
- 689 - The alternative formulations for BAMF, BAMF-AR1 and BAMF-GS, did not show a significant
690 improvement with respect to BAMF.

691 With all that, profile sparsity has been shown to substantially enhance the accuracy of source apportionment
692 analyses, improving the separation of the chemical composition of sources. The BAMF+HS model succeeds in
693 incorporating this property in profile fingerprints, especially in filter-based datasets. Using BAMF+HS in such
694 datasets, the solutions reflect the sparsity of filter-based chemical profiles, hence, this newly introduced method
695 is encouraged when source fingerprints are expected to be substantially sparse. However, for ACSM-like datasets,
696 the sparsity is not fully achieved due to converge issues, although the quality of the solution is not substantially
697 deprecated with respect to BAMF. With the aim of improving further source apportionment techniques, future
698 research should be directed to enhance the robustness and generalisability of the BAMF+HS model across diverse
699 data types. Moreover, continued exploration of the underlying properties of solution spaces (such as profile
700 sparsity, time series autocorrelation) may provide valuable insights into disentangling complex source
701 contributions through receptor modelling. In this regard, the Bayesian source apportionment framework offers a
702 particularly suitable foundation, allowing for the integration of prior knowledge and uncertainty quantification in
703 the inference process.

704 **Code and data availability**

705 The models and datasets can be found at <https://github.com/martavia0/BAMF-horseshoe.git>

706 **Author contribution**

707 MV: Conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project
708 administration, resources, software, validation, visualisation, writing (original draft preparation). YH: Formal
709 analysis, investigation, software; JD: investigation, resources, software, validation. MM: Data curation. AR: Data
710 curation, formal analysis, methodology, investigation, resources, software. JJ: Data curation. SKG: Data curation.
711 J-LJ: Data curation; VNTD: Data curation. GU: Data curation. GM: conceptualisation, funding acquisition,
712 investigation, supervision, validation. KRD: Conceptualisation, data curation, formal analysis, funding
713 acquisition, investigation, methodology, supervision, validation. All co-authors participated in the revision and
714 edition of the manuscript.

715

716 **Competing interests**

717 The authors declare that they have no conflict of interest.

718

719 **Disclaimer**

720

721 Co - funded by the European Union. Views and opinions expressed are however those of the author(s) only and
722 do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the
723 European Union nor the granting authority can be held responsible for them.

724 **Acknowledgements**

725 This publication is co- funded by/ has received funding from/ the European Union’s Horizon Europe research and
726 innovation program under the Marie Skłodowska-Curie COFUND Postdoctoral Programme grant agreement
727 No.101081355- SMASH and by/from /the Republic of Slovenia and the European Union from the European
728 Regional Development Fund. Work was partly supported by ARIS program P1-0385. All the chemical
729 measurements on the PM10 and PM2.5 Swiss samples used in this study were performed on the Air O Sol
730 analytical plateau at IGE (France). K.R.D. acknowledges support by the SNSF Ambizione grant PZPGP2_201992.
731 The authors thank Jimeng Wu and Imad El Haddad for their input related to autocorrelation.

732

733 **References**

734 Andersen, M. R., Winther, O., & Hansen, L. K. (2014). Bayesian inference for structured spike and slab
735 priors. *Advances in Neural Information Processing Systems*, 27.

736 Belis, C. A., Karagulian, F., Larsen, B. R., & Hopke, P. K. (2013). Critical review and meta-analysis of ambient
737 particulate matter source apportionment using receptor models in Europe. *Atmospheric Environment*, 69, 94-108.

738 Belis, C. A., Larsen, B. R., Amato, F., El Haddad, I., Favez, O., Harrison, R. M., ... & Viana, M. (2014). European
739 guide on air pollution source apportionment with receptor models. JRC reference reports EUR26080 EN.

740 Belis, C., Pernigotti, D., Karagulian, F., Pirovano, G., Larsen, B., Gerboles, M., and Hopke, P.: A new
741 methodology to assess the performance and uncertainty of source apportionment models in intercomparison
742 exercises, *Atmospheric Environment*, 119, 35–44, 2015.

743 Brinkman, G., Vance, G., Hannigan, M. P., and Milford, J. B.: Use of synthetic data to evaluate positive matrix
744 factorization as a source apportionment tool for PM2. 5 exposure data, *Environmental science & technology*, 40,
745 1892–1901, 2006.

746 Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan:
747 A probabilistic programming language. *Journal of statistical software*, 76, 1-32.

748 Crippa, M., DeCarlo, P. F., Slowik, J. G., Mohr, C., Heringa, M. F., Chirico, R., ... & Baltensperger, U. (2013).
749 Wintertime aerosol chemical composition and source apportionment of the organic fraction in the metropolitan
750 area of Paris. *Atmospheric Chemistry and Physics*, 13(2), 961-981.

751 Dai, T., Dai, Q., Yin, J., Chen, J., Liu, B., Bi, X., ... & Feng, Y. (2024). Spatial source apportionment of airborne
752 coarse particulate matter using PMF-Bayesian receptor model. *Science of The Total Environment*, 917, 170235.

753 Daellenbach, K. R., Uzu, G., Jiang, J., Cassagnes, L. E., Leni, Z., Vlachou, A., ... & Prévôt, A. S. (2020). Sources
754 of particulate-matter air pollution and its oxidative potential in Europe. *Nature*, 587(7834), 414-419.

755 Daellenbach, K. R., Manousakas, M., Jiang, J., Cui, T., Chen, Y., El Haddad, I., ... & Prévôt, A. S. H. (2023).
756 Organic aerosol sources in the Milan metropolitan area—Receptor modelling based on field observations and air
757 quality modelling. *Atmospheric Environment*, 307, 119799.

758 Dinh, V. N. T., Uzu, G., Dominutti, P., Sauvage, S., Elazzouzi, R., Darfeuil, S., Voiron, C., Samaké, A., Zhang,
759 S., Socquet, S., Favez, O., and Jaffrezo, J.-L.: Toolbox for accurate estimation and validation of PMF solutions in
760 PM source apportionment, *EGUsphere* [preprint], <https://doi.org/10.5194/egusphere-2025-1968>, 2025.

761 Elser, M., Huang, R. J., Wolf, R., Slowik, J. G., Wang, Q., Canonaco, F., ... & Prévôt, A. S. (2016). New insights
762 into PM 2.5 chemical composition and sources in two major cities in China during extreme haze events using
763 aerosol mass spectrometry. *Atmospheric Chemistry and Physics*, 16(5), 3207-3225.

764 Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical*
765 *Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>

766 Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: *Bayesian data analysis*, 3rd
767 edn., CRC Press, ISBN 9781439898208, 2014.

768 Gini, Corrado (1936). "On the Measure of Concentration with Special Reference to Income and Statistics",
769 Colorado College Publication, General Series No. 208, 73–79

770 Govan, E., Jackson, A. L., Inger, R., Bearhop, S., & Parnell, A. C. (2023). *simmr*: A package for fitting stable
771 isotope mixing models in R. *arXiv preprint arXiv:2306.07817*.

772 Grange, S. K., Fischer, A., Zellweger, C., Alastuey, A., Querol, X., Jaffrezo, J. L., ... & Hueglin, C. (2021).
773 Switzerland's PM10 and PM2.5 environmental increments show the importance of non-exhaust emissions.
774 *Atmospheric environment: X*, 12, 100145.

775 Hoffman, M. D. and Gelman, A.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte
776 Carlo, *J. Mach. Learn. Res.*, 15, 1593–1623, 2014.

777 Jiang, J., Aksoyoglu, S., El-Haddad, I., Ciarelli, G., Denier van der Gon, H. A., Canonaco, F., ... & Prévôt, A. S.
778 (2019). Sources of organic aerosols in Europe: A modelling study using CAMx with modified volatility basis set
779 scheme. *Atmospheric Chemistry and Physics Discussions*, 2019, 1-35.

780 Keuken, M. P., Moerman, M., Voogt, M., Blom, M., Weijers, E. P., Röckmann, T., & Dusek, U. (2013). Source
781 contributions to PM2.5 and PM10 at an urban background and a street location. *Atmospheric Environment*, 71,
782 26-35.

783 Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-
784 2), 83-97.

785 Lingwall, J. W., & Christensen, W. F. (2007). Pollution source apportionment using a priori information and
786 positive matrix factorization. *Chemometrics and Intelligent Laboratory Systems*, 87(2), 281-294.

787 Liu, D. C. and Nocedal, J.: On the Limited Memory BFGS Method for Large Scale Optimization, *Math. Program.*,
788 45, 503–528, <https://doi.org/10.1007/BF01589116>, 1989.

789 Manousakas, M. I., Rausch, J., Jaramillo Vogel, D., Schneider-Beltran, K., Alastuey, A., Jaffrezo, J. L., ... &
790 Dällenbach, K. R. Comparison of PM Source Profiles Identified by Different Techniques and the Potential of
791 Utilizing Single-Particle Analysis Data in Source Apportionment. Available at SSRN 5323830.

792 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state
793 calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.
794 <https://doi.org/10.1063/1.1699114>

795 Oh, M. S., & Park, C. K. (2022). Regional source apportionment of PM_{2.5} in Seoul using Bayesian multivariate
796 receptor model. *Journal of Applied Statistics*, 49(3), 738-751.

797 Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal
798 utilization of error estimates of data values. *Environmetrics*, 5(2), 111-126.

799 Paatero, P. (1999). The multilinear engine—a table-driven, least squares program for solving multilinear
800 problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*,
801 8(4), 854-888.

802 Park, E. S., Guttorp, P., & Henry, R. C. (2001). Multivariate receptor modeling for temporally correlated data by
803 using MCMC. *Journal of the American Statistical Association*, 96(456), 1171-1183.

804 Park, E. S., Spiegelman, C. H., & Henry, R. C. (2002). Bilinear estimation of pollution source profiles and amounts
805 by using multivariate receptor models. *Environmetrics*, 13(7), 775-798.

806 Park, E. S., Lee, E. K., & Oh, M. S. (2021). Bayesian multivariate receptor modeling software: BNFA and
807 bayesMRM. *Chemometrics and Intelligent Laboratory Systems*, 211, 104280.

808 Piironen, J. and Vehtari, A.: Sparsity information and regularization in the horseshoe and other shrinkage priors,
809 *Electronic Journal of Statistics*, 11, 5018–5051, <https://doi.org/10.1214/17-EJS1337SI>, 2017.

810 Rasmussen, M. A., & Bro, R. (2012). A tutorial on the Lasso approach to sparse modeling. *Chemometrics and*
811 *Intelligent Laboratory Systems*, 119, 21-31.

812 Pope III, C. A., & Dockery, D. W. (1999). Epidemiology of particle effects. In *Air pollution and health* (pp. 673-
813 705). Academic Press.

814 Rusanen, A., Bjorklund, A., Manousakas, M. I., Jiang, J., Kulmala, M. T., Puolamaki, K., and Daellenbach, K.
815 R.: A novel probabilistic source apportionment approach: Bayesian auto-correlated matrix factorization,
816 *Atmospheric Measurement Techniques*, 17, 1251–1277, <https://doi.org/10.5194/amt-17-1251-2024>, 2024.

817 Sage, A. M. (2007). Evolving mass spectra of the oxidized component of organic aerosol mass spectrometer
818 analysis of aged diesel emissions. *Atmos. Chem. Phys. Discuss.*, 7, 10065-10096.

819 STAN Development Team. (2025). *Stan user's guide* (Version 2.36). <https://mc-stan.org/docs/stan-users-guide/>.
820 Accessed July 2025.

821 Tobler, A. K., Skiba, A., Canonaco, F., Močnik, G., Rai, P., Chen, G., ... & Prevot, A. S. (2021). Characterization
822 of non-refractory (NR) PM₁ and source apportionment of organic aerosol in Kraków, Poland. *Atmospheric*
823 *chemistry and physics*, 21(19), 14893-14906.

824 Tibshirani, R., & Wasserman, L. (2015). Sparsity and the lasso. *Statistical machine learning*, 1-15.

825 Ulbrich, I. M., Handschy, A., Lechner, M., and Jimenez, J. L.: AMS Spectral Database,
826 <http://cires.colorado.edu/jimenez-group/AMSsd/>, accessed: 2025-05-19, n.d.

827 Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., & Jimenez, J. L. (2009). Interpretation of organic
828 components from Positive Matrix Factorization of aerosol mass spectrometric data. *Atmospheric Chemistry and*
829 *Physics*, 9(9), 2891-2918.

830 Upadhyay, A., Jiang, J., Cheng, Y., Vasilakos, P., Chen, Y., Banos, D. T., ... & El-Haddad, I. (2025). High-
831 resolution modelling of particulate matter chemical composition over Europe: brake wear pollution. *Environment*
832 *International*, 109615.

833 Via, M., Minguillón, M. C., Reche, C., Querol, X., & Alastuey, A. (2021). Increase in secondary organic aerosol
834 in an urban environment. *Atmospheric chemistry and physics*, 21(10), 8323-8339.

835 Via, M., Chen, G., Canonaco, F., Daellenbach, K. R., Chazean, B., Chebaicheb, H., Jiang, J., Keernik, H., Lin,
836 C., Marchand, N., et al.: Rolling vs. seasonal PMF: real-world multi-site and synthetic dataset comparison,
837 Atmospheric measurement techniques, 15, 5479–5495, 2022.

838 Viana, M., Kuhlbusch, T. A., Querol, X., Alastuey, A., Harrison, R. M., Hopke, P. K., ... & Hitzenberger, R.
839 (2008). Source apportionment of particulate matter in Europe: a review of methods and results. Journal of aerosol
840 science, 39(10), 827-849.

841 Yang, Y., Ruan, Z., Wang, X., Yang, Y., Mason, T. G., Lin, H., & Tian, L. (2019). Short-term and long-term
842 exposures to fine particulate matter constituents and health: A systematic review and meta-
843 analysis. Environmental pollution, 247, 874-882.

844 Zhang, Y., Albinet, A., Petit, J. E., Jacob, V., Chevrier, F., Gille, G., ... & Favez, O. (2020). Substantial brown
845 carbon emissions from wintertime residential wood burning over France. Science of the Total Environment, 743,
846 140752.

847 Zhang, Y., Ma, X., Tang, A., Fang, Y., Misselbrook, T., & Liu, X. (2023). Source Apportionment of Atmospheric
848 Ammonia at 16 Sites in China Using a Bayesian Isotope Mixing Model Based on $\delta^{15}\text{N-NH}_x$
849 Signatures. Environmental Science & Technology, 57(16), 6599-6608.

850 Zheng, Y., Xue, T., Zhao, H., & Lei, Y. (2022). Increasing life expectancy in China by achieving its 2025 air
851 quality target. Environmental science and ecotechnology, 12, 100203.

852 **Figures**

853 **Table 1. Models used in the current study and their priors on the G, F matrices.**

854

Model	G priors	F priors
BMF	None	None
BMF+HS	None	Regularised horseshoe
BAMF	Rusanen et al. (2024)	None
BAMF+HS	Rusanen et al. (2024)	Regularised horseshoe
BAMF-AR1	AR(1)	None
BAMF-AR1+HS	AR(1)	Regularised horseshoe
BAMF-GS	Rusanen et al. (2024)	None
PMF	None	None
CMB	Fixed a-priori.	None
CMB+HS	Fixed a-priori.	Regularised horseshoe

855

856 **Table 2. Profile sparsity metrics for the truth of synthetic datasets.**

857

Dataset	Factor	F Gini	% zeros

Chemically-sparse synthetic toy dataset	Factor 1		0.67	75.0
	Factor 2		0.67	75.0
	Factor 3		0.5	25.0
Chemically-sparse synthetic offline dataset	Dust		0.74	37.5
	Traffic		0.86	12.5
	Salt		0.78	37.5
	Coarse biological		0.88	25.0
Less chemically-sparse synthetic online ACSM datasets	HOA	Krakow	0.74	5.0
		Milan	0.67	0.0
		Paris	0.68	0.0
		Zurich	0.68	0.0
	BBOA	Krakow	0.47	1.2
		Milan	0.72	0.0
		Paris	0.74	0.0
		Zurich	0.58	0.0
	SOA _{bio}	Krakow	0.52	0.0
		Milan	0.50	0.0
		Paris	0.50	0.0
		Zurich	0.67	0.0
	SOA _{BB}	Krakow	0.55	0.0
		Milan	0.53	0.0
		Paris	0.53	0.0

		Zurich	0.73	0.0
	SOA _{TR}	Krakow	0.45	0.0
		Milan	0.42	0.0
		Paris	0.46	0.0
		Zurich	0.60	0.0

858

859 **Table 3. Toy experiment statistics of (a) Factorisation performance. (b) Comparison to truth. Green**
860 **sequential colorscales represent variables whose larger value leads to a better performance and the blue-**
861 **to-red divergent colorscales (centered at 1, in white) represent G/G_0 divergence with respect to 1. Red bars**
862 **in (a) depict deviations from the ideal 0 value.**

863

(a)

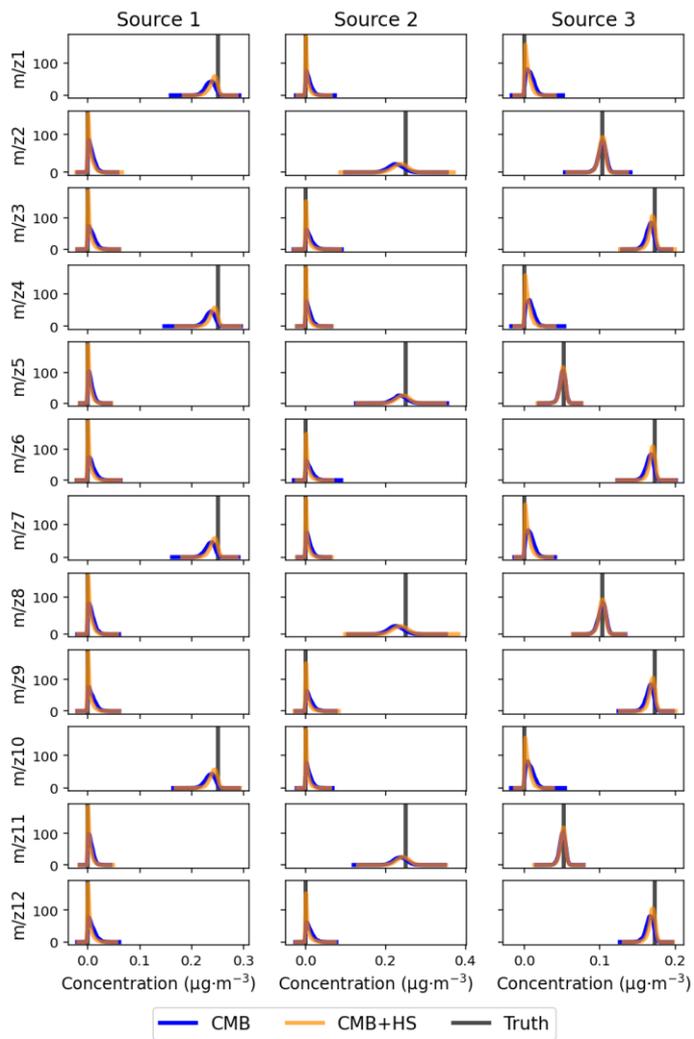
Model	Factorisation		
	R ²	Median(Z-X /sigma)	Max(Z-X /sigma)
CMB	0.9985	0.2542	0.5755
CMB+HS	0.9996	0.2648	0.4774
PMF	0.9979	0.2658	0.5847
BMF	0.9650	0.2125	1.2917
BMF+HS	0.9689	0.2107	1.2863
BAMF	0.9818	0.1222	0.7247
BAMF+HS	0.9820	0.1208	0.7180
BAMF-ARI	0.9806	0.0947	1.0257
BAMF-ARI+HS	0.9790	0.1171	1.0297
BMF-GS	0.9657	0.2024	1.3031
BAMF-GS	0.9818	0.1576	0.9082

(b)

Model	Sources	G				F		
		G/G ₀	R ²	ρ	R ²	Gini	Gini ratio	Zeros sum
CMB	Source1	1.00	1.00	0.82	1.00	0.61	0.91	0.06
	Source2	1.00	1.00	0.82	1.00	0.60	0.90	0.08
	Source3	1.00	1.00	0.96	1.00	0.44	0.92	0.03
CMB+HS	Source1	1.00	1.00	0.82	1.00	0.63	0.95	0.03
	Source2	1.00	1.00	0.82	1.00	0.63	0.95	0.04
	Source3	1.00	1.00	0.96	1.00	0.47	0.96	0.02
PMF	Source1	1.00	0.96	0.82	0.95	0.35	0.52	0.36
	Source2	3.29	0.97	0.82	0.88	0.30	0.45	0.42
	Source3	0.65	0.74	0.79	0.81	0.31	0.64	0.2
BMF	Source1	1.08	0.87	0.82	0.76	0.25	0.37	0.48
	Source2	4.10	0.84	0.51	0.43	0.16	0.23	0.58
	Source3	0.60	0.46	0.79	0.80	0.21	0.44	0.25
BMF+HS	Source1	1.11	0.88	0.82	0.77	0.25	0.38	0.48
	Source2	3.92	0.85	0.82	0.50	0.15	0.23	0.57
	Source3	0.60	0.44	0.79	0.84	0.22	0.46	0.23
BAMF	Source1	1.66	0.98	0.82	0.78	0.27	0.41	0.46
	Source2	1.99	0.95	0.82	0.97	0.33	0.50	0.37
	Source3	0.54	0.50	0.96	1.00	0.38	0.79	0.08
BAMF+HS	Source1	1.96	0.98	0.82	0.76	0.27	0.40	0.47
	Source2	1.28	0.95	0.82	0.99	0.58	0.86	0.11
	Source3	0.51	0.48	0.96	1.00	0.44	0.93	0.02
BAMF-ARI	Source1	1.70	0.98	0.82	0.79	0.27	0.41	0.46
	Source2	3.05	0.92	0.82	0.89	0.37	0.55	0.36
	Source3	0.42	0.50	0.96	0.94	0.43	0.90	0.08
BAMF-ARI+HS	Source1	1.92	0.99	0.82	0.77	0.26	0.39	0.46
	Source2	2.58	0.89	0.82	0.90	0.48	0.72	0.36
	Source3	0.38	0.49	0.96	0.97	0.49	1.03	0.08
BMF-GS	Source1	0.88	0.92	0.82	0.63	0.21	0.31	0.52
	Source2	1.12	0.86	0.82	0.71	0.20	0.30	0.53
	Source3	1.02	0.29	0.79	0.91	0.22	0.46	0.22
BAMF-GS	Source1	0.89	0.94	0.82	0.66	0.21	0.32	0.53
	Source2	1.12	0.95	0.82	0.70	0.19	0.28	0.53
	Source3	1.02	0.36	0.79	0.90	0.22	0.46	0.22
BAMF-Lasso	Source1	1.76	0.98	0.82	0.76	0.27	0.40	0.47
	Source2	2.02	0.95	0.82	0.95	0.34	0.50	0.37
	Source3	0.49	0.47	0.96	1.00	0.41	0.86	0.05
BAMF-Spike-Slab	Source1	1.57	0.97	0.82	0.77	0.26	0.40	0.47
	Source2	1.78	0.95	0.82	0.99	0.34	0.52	0.34
	Source3	0.62	0.50	0.96	0.98	0.33	0.69	0.12

864

865

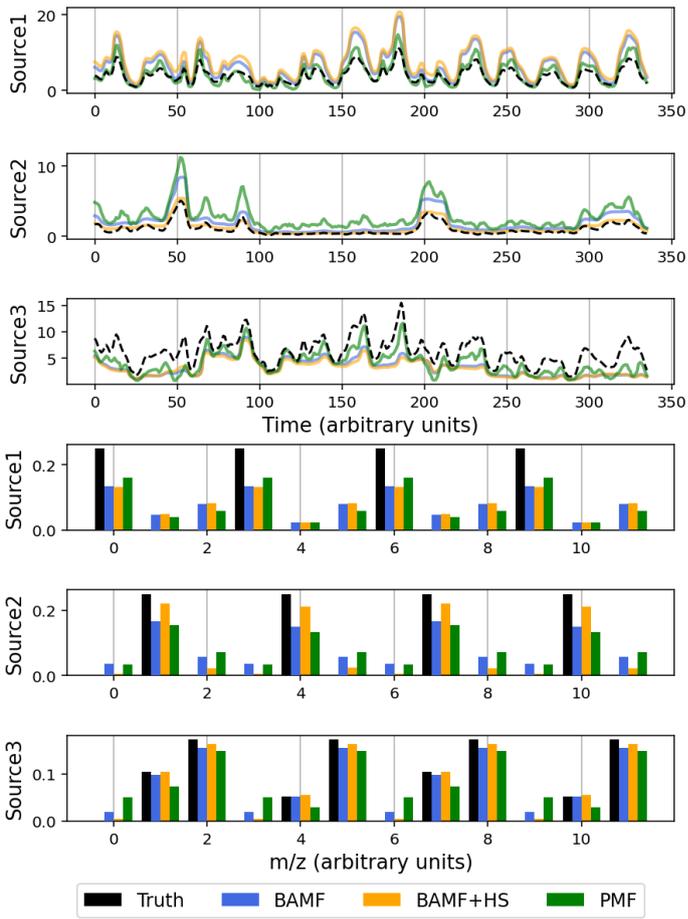


866

867 **Figure 1. Distributions for the mass concentrations of all measured variables (m/z) for both CMB and CMB+HS (solid**
 868 **lines) compared to truth (markers).**

869

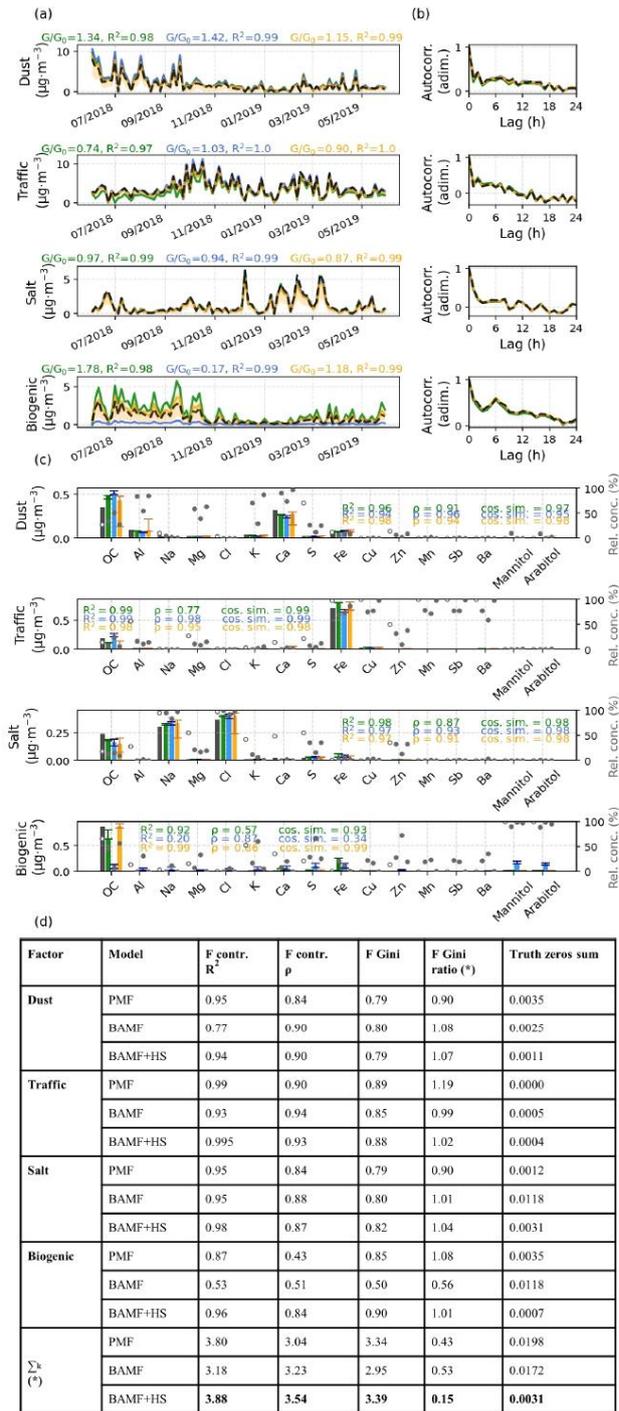
870



871

872 **Figure 2. Source apportionment results for the toy dataset obtained using PMF, BAMF, and BAMF+HS, compared**
 873 **against the true solution (black bars). (a) Factor time series. (b) Factor profiles.**

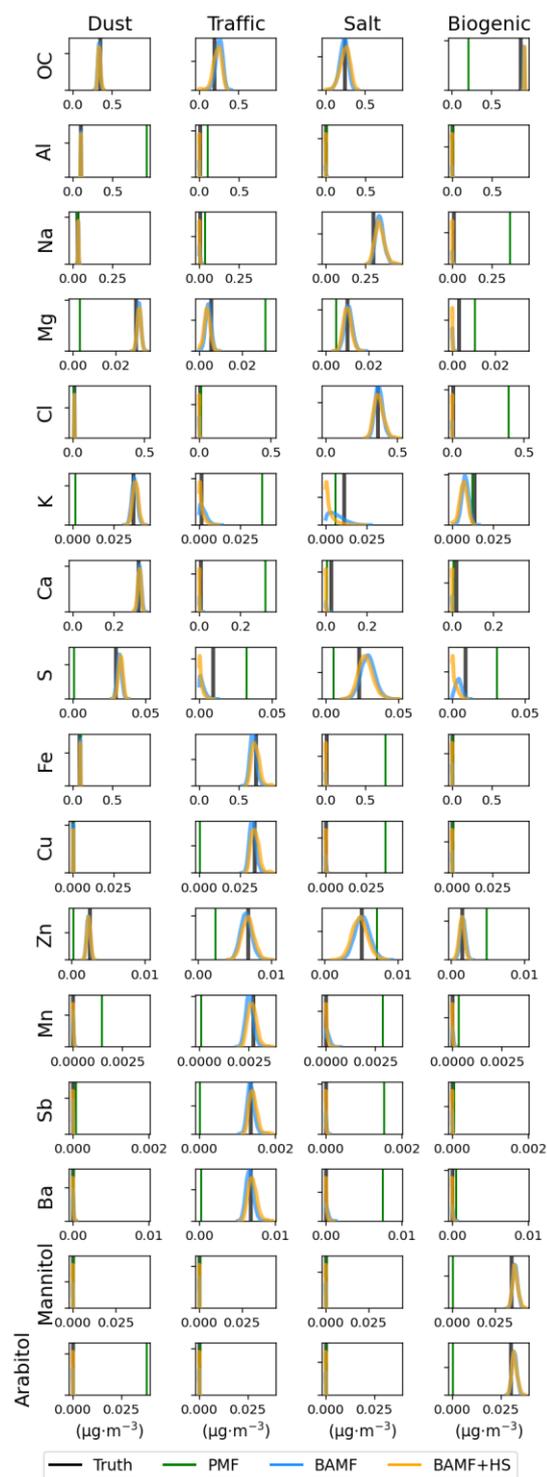
874



875

876 **Figure 3. Synthetic offline dataset source apportionment results for PMF, BAMF, and BAMF+HS models. (a) Time**
 877 **Series. (b) Autocorrelation. (c) Profiles. (d) Table with additional metrics for comparison to truth. Bold numbers reflect**
 878 **the highest value amongst models. F contr. represents here the percentage of each factor in a given species. The sum**
 879 **row reflects the overall performance of the model for all sources for each statistic metric except for the ones marked**
 880 **with (*), in which the difference to 1 in absolute value is summed up.**

881



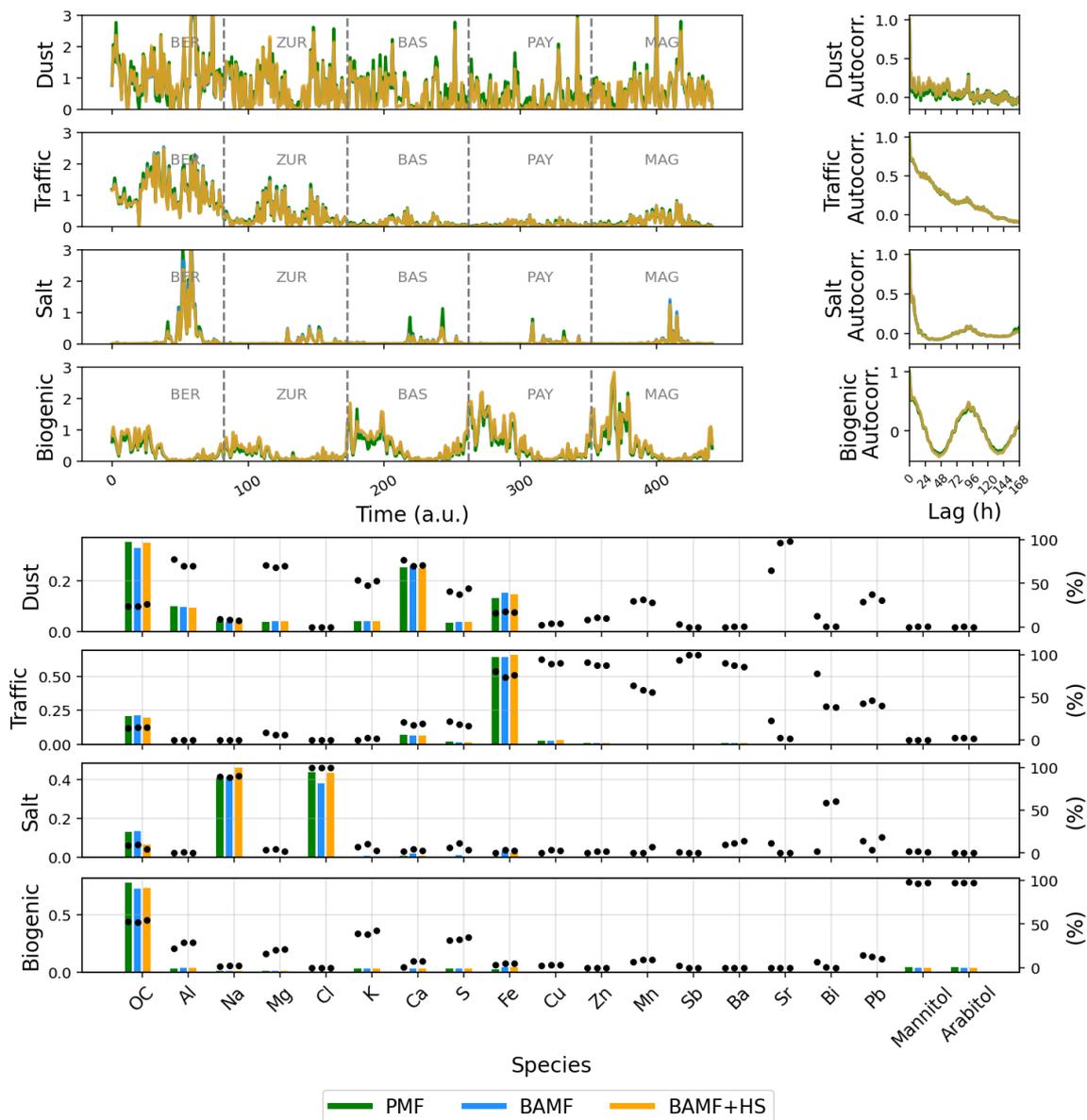
882

883

884 **Figure 4. Profile components distribution for PMF, BAMF, BAMF+HS (colored lines) in comparison to the**
 885 **truth (black lines) on the real-world filters dataset. Rows represent the species of the source apportionment**
 886 **and columns represent sources.**

887

888



889

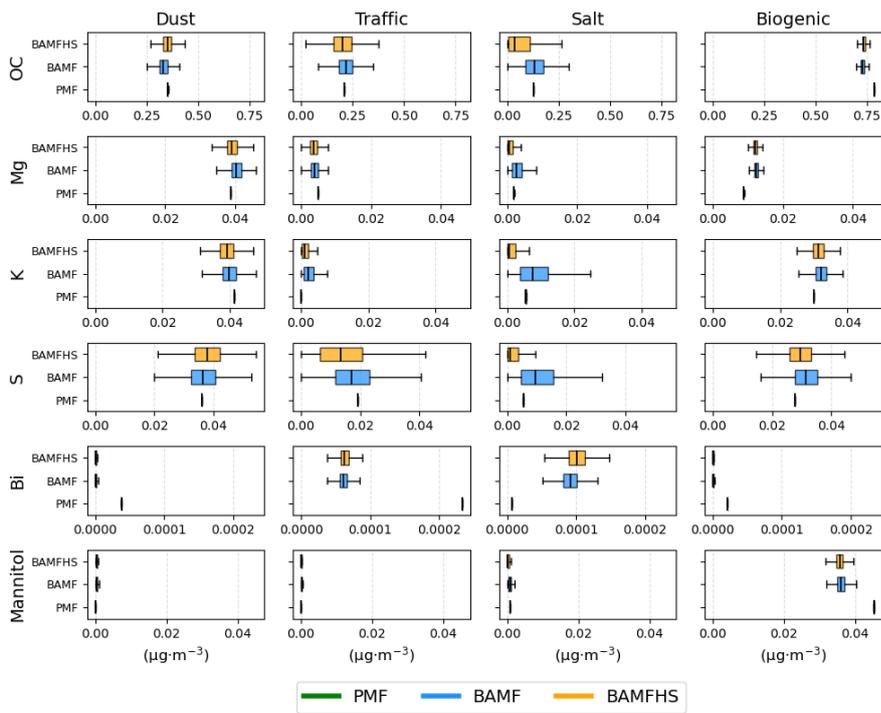
890 **Figure 5. Comparison of PMF, BAMF, BAMF+HS for the real-world filters dataset. From left to right and**
 891 **top to bottom: time series, autocorrelation, and profile plots. The dots in the profiles (right axis) show the**
 892 **contribution of each species to the source.**

893 **Table 4. Offline real-world dataset reconstruction and sparsity statistics. Bold numbers reflect the highest**
 894 **value amongst models.**

Model	$R^2(Z, X)$	Median $ X-Z /\sigma$	Median $ X-Z /\sigma$	Factor	F Gini
PMF	0.68	0.77	10.52	Dust	0.77
				Traffic	0.87
				Salt	0.86
				Biogenic	0.87

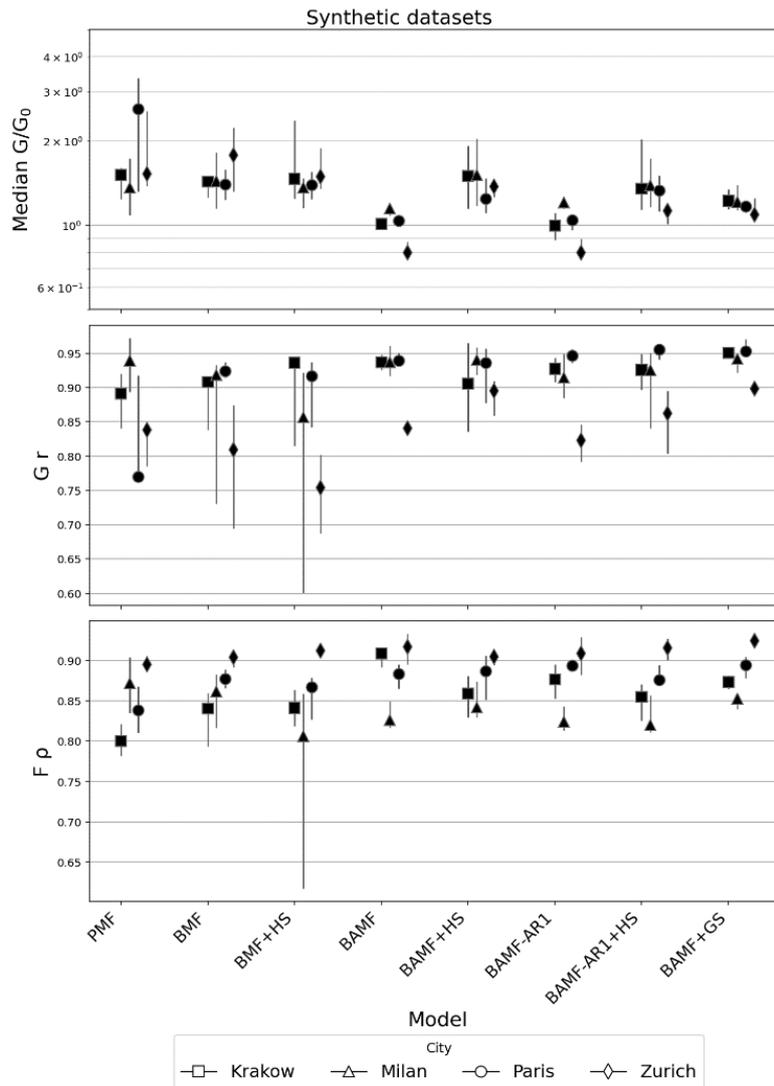
BAMF	0.67	0.75	11.13	Dust	0.76
				Traffic	0.87
				Salt	0.84
				Biogenic	0.83
BAMF+HS	0.67	0.75	11.12	Dust	0.77
				Traffic	0.87
				Salt	0.87
				Biogenic	0.83

895
896
897
898
899



900
901
902
903

Figure 6. Boxplot distributions of individual profile components derived from PMF, BAMF, and BAMF+HS analyses for the real-world filter dataset. A complete comparison of all profiles is presented in Figure S9.



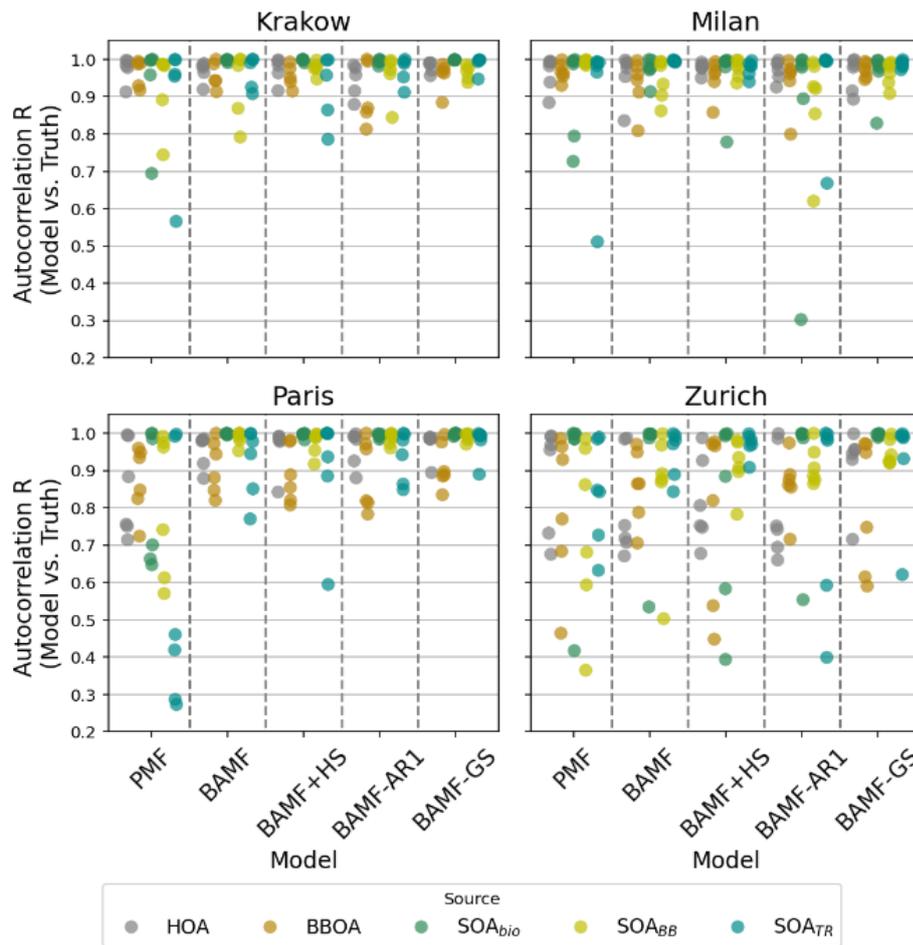
904

905

906

907

Figure 7. European cities synthetic datasets summary statistics; from top to bottom, median ratio time series with truth (G/G_0), Pearson correlation coefficient of G with truth ($G r$), Spearman correlation coefficient of F with truth ($F \rho$). The axis of the G/G_0 plot is in logarithmic scale.

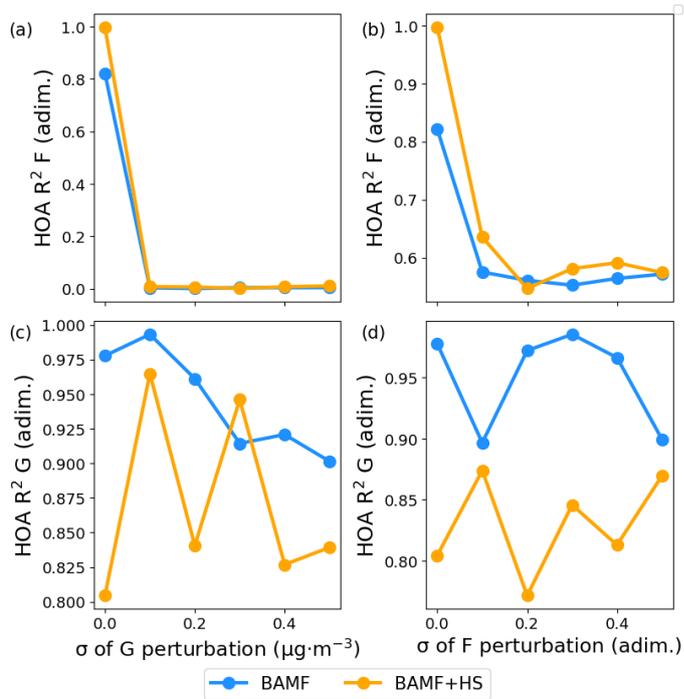


908

909 **Figure 8. Pearson correlation of the autocorrelations of model solutions with the truth for all factors and all cities.**

910

911



912

913 **Figure 9. Squared Pearson coefficient of F, G matrix with original truth F, G matrices of the BAMF, BAMF+HS models**
 914 **with the degrees of perturbation in F and G.**

915