Response to Reviewer #1

In this paper the authors present a very detailed and thorough anlysis of sparse methods for selection of relevant absorption bands in IR. The paper is well presented and the analysis is very detailed. I only have a few general comments regarding the general accuracy of methods according to reliability of the sampled data.

We thank the reviewer for the positive comments.

1. Page 2 line 24 onwards. Is the justification for sparse methods, in this sense, more with regards the end-point accuracy or prohibitive cost of other methods?

We would like to use this opportunity to discuss our separate motivations for 1) using sparse methods, and 2) FT-IR more generally.

The motivation for implementing sparse methods is to improve end-point accuracy and interpretation of aerosol composition by FT-IR. Uninformative variables (i.e., wavenumbers) can add variance, or in some instances, bias to the predictions. In this work, the advantage with respect to end-point accuracy is not very clear, as the original full wavenumber models were already quite accurate. However, elimination of uninformative variables helps drive our interpretation regarding the most important absorption bands necessary for predicting TOR OC and EC, as well as FG, which allows us to make associations between carbonaceous aerosol and their molecular structure.

Our focus on developing quantiative algorithms FT-IR more generally is motivated for its potential to provide 1) a reduction in cost of aerosol monitoring, and 2) the additional chemical resolution (via functional groups) not permitted by other instruments. FT-IR analysis of each PTFE sample requires only a few minutes and can be integrated into existing analysis toolchains without the need for a separate sampling line for aerosol collection on quartz fiber filters (specifically for thermal optical analysis). In addition, absorption by vibrational modes observed by FT-IR can inform us of aerosol molecular structure not provided by organic carbon analysis, or even common mass spectrometry techniques. We anticipate that FG information harvested from such measurements can be made useful for systematic model-measurement evaluation (e.g., Ruggeri et al., 2016).

As the first point is the focus of (Ruthenburg et al., 2014; Dillner and Takahama, 2015a,b; Reggente et al., 2016), we include this addition to the Abstract:

"Various vibrational modes present in molecular mixtures of laboratory and atmospheric aerosols give rise to complex Fourier Transform Infrared (FT-IR) absorption spectra. Such spectra can be chemically-informative, but often require sophisticated algorithms for quantitative characterization of aerosol composition. Naïve statistical calibration models developed for quantification employ the full suite of wavenumbers available from a set of spectra, leading to loss of mechanistic interpretation between chemical composition and the resulting changes in absorption patterns that underpin their predictive capability. Using sparse representations of the same set of spectra, alternative calibration models can be built in which only a select group of absorption bands are used to make quantitative prediction of various aerosol properties. Such models are desirable as they allow us to make association

of predicted properties with their underlying molecular structure."

2. In section 2.1.1 the laboratory and ambient samples are discussed. It would be nice to see a brief consideration of instrument variability and how this might impact of the performance of your evaluation. For example, is there an underlying assumption that there is no 'drift' in instrument response function across all datasets used? I wonder whether this might lead to the variability discussed in section 3.2 onwards? Perhaps I have misunderstood this, but is agreement between instruments proven?

The agreement between instruments for this dataset has indeed been established by Dillner and Takahama (2015a,b). We have included this statement in the Introduction section (1):

"We remark that while the latter two carbonaceous substances [TOR OC and EC] are presumably are composed of a complex combination of molecules, statistical calibration models using FT-IR spectra have been demonstrated to accurately predict TOR-equivalent OC and EC within measurement precision of collocated samples reported by the TOR method (Dillner and Takahama, 2015a,b)."

and in Methods section (2.3.1):

"Full wavenumber calibration models for TOR OC and EC developed using this protocol have demonstrated ability to capture overall variations in concentrations which span a wide range of PM composition and environmental conditions (Dillner and Takahama, 2015a,b)."

Regarding the specific issue of instrument drift, we do not attribute discrepancies between predictions to changes in instrument response for this work. The TOR measurements follow established protocols for QA/QC (http://vista.cira.colostate.edu/improve/data/QA_QC/qa_qc_Branch.htm) in which samples analyzed during periods of excessive drift are reanalyzed. We are preparing separate documentation of performance evaluation for FT-IR, but repeated analysis of check standard samples indicate less than 2–5% drift in signal response over 1.5 years.

3. Towards the end of page 6 the authors discuss the rationale for choice of variable selection method. In section 2.3.1, how is the test fair in the sense of being able to cover a broad range of functionality expected in amb[i]ent samples? In other words, would it be possible to visualise where your test and training set 'sit' on a general 2D basis sets or even simple Ven Krevelen diagrams? Does this even matter?

We thank the reviewer for this question.

The range in functionality expected in ambient samples does guide our choice of calibration set samples and the final variables selected (though not the variable selection procedure, as this is statistically determined).

Apart from the chamber studies of Chhabra et al. (2011), we have not evaluated the O/C and H/C ratio characterized by FT-IR measurements in detail. This is an area of current work. However, the range in OM/OC ratio (which can be closely associated with O/C; Aiken et al., 2008) over the IMPROVE 2011 samples is described and presented by Ruthenburg et al. (2014). Selection of laboratory standards used for prediction of FG is not selected on the basis of OM/OC, however, but on the multidimensional space of the FGs such that overlapping absorption bands can be taken into account in the model development. One of the reasons for the larger range in predicted FG abundance among models is due to the fact that the laboratory standards and ambient samples are different, and constraining the model development and selection under these circumstances is an area

of current research.

For TOR OC analysis, Dillner and Takahama (2015a) explored the influence of aerosol composition on calibration set design. Aerosol composition was parameterized as OM/OC ratio (derived from estimates of Ruthenburg et al., 2014) and ammonium/OC ratio, and we found, not unexpectedly, that capability for prediction was dependent on the range of concentration and similarity composition between the calibration (which comprises training and validation) and test set samples. Similar conclusions for EC are drawn (Dillner and Takahama, 2015b). Therefore, the calibration and test set samples are selected to span the same composition space (e.g., as might be represented by van Krevlin diagram) by design. The variable (wavenumbers) selected for TOR OC and EC are specific to the composition of PM sampled in these seven sites for 2011, though Reggente et al. (2016) suggests the analysis can be extended for the same sites to 2013. More importantly, we feel that the main contribution of this manuscript is to introduce this technique and open up a new avenue of research, in which we can associate vibrational modes and molecular structure with aerosol of different composition (e.g., targeting specific regions in the van Krevlin space) characterized by collocated measurements; not limited to TOR. Such work is currently underway in our group.

We have included the following statement in the Methods section (2.1.1):

"The OM/OC ratio estimated in ambient samples span a range of 1.46 and 2.01 between the 10th and 90th percentiles, with a median ratio of 1.69 (Ruthenburg et al., 2014)."

and the following statement in the Methods section (2.3.1):

"The calibration and test sets are constructed identically to the most accurate class of full wavenumber models described previously (Ruthenburg et al., 2014; Takahama and Dillner, 2015; Dillner and Takahama, 2015a,b)."

and the following statement in the Conclusion section (4):

"Sparse calibration models "localized" (in the statistical sense) by spectral features or external variables can be used to identify key FGs used for prediction of TOR measurements at various sites (e.g., Reggente et al., 2016), and aid construction of calibration sets suitable for prediction of individual or groups of samples (e.g., stratified by environmental conditions or chemical composition). This work provides a demonstration of how molecular structure can be associated with other quantifiable metrics of complex PM to which spectral features from FT-IR can be correlated."

4. On this point, there seems to be complicating factors in sample collection from ambient campaigns that might degrade the performance of any 'tuned' method. For example, could it be that whether any semi-volatile loss, for example, occured during sample preparation might lead to a functional group dependency of variable selection? It might be prudent to assess such effects with a consideration of other factors such as RH, Temp, provenance of sample. Based on the high mass loadings displayed in figure 3, this might not yet be apparent but its worthwhile considering. I guess I'm wondering whether, aside from variable selection refining a fixed instrument response, the algorithms are also picking up 'noise' from external factors.

We thank the reviewer for raising this important point. Investigation of artifacts, and particularly semi-volatile species, is an active area of investigation for us and will be addressed in a series of future studies. However, for the purposes of this manuscript, we note that Ruthenburg et al. (2014); Takahama and Dillner (2015); Dillner and Takahama (2015a,b)

have shown that a single set of calibration models can make accurate predictions for each of the response variables studied. Predicted concentrations of individual FGs have not been independently validated, but have been evaluated to the extent possible using agreement with TOR OC and trends in OM/OC ratios.

Therefore, to a first order we believe that we have been able to capture the varying range in aerosol composition on these filters due to effects of RH, temperature, semivolatile losses with our base case (full wavenumber) models. As the TOR OC and EC estimates are calibrated to collocated measurements, there is a question whether sampling artifacts may be different. Quartz fiber filters have vapor adsorption artifacts, but have been corrected in a "mean" way for OC (Dillner and Takahama, 2015a).

We have revised Methods section (2.1.1):

"Monthly median values of OC loadings in blank samples are subtracted from ambient TOR OC loadings to account for the gas phase adsorption artifact by quartz fiber filters (Dillner and Takahama, 2015a)."

and Methods section (2.3.1):

"Full wavenumber calibration models for TOR OC and EC developed using this protocol have demonstrated ability to capture overall variations in concentrations which span a wide range of PM composition and environmental conditions (Dillner and Takahama, 2015a,b)."

5. [Minor comment] Page 8, line 18: I think an if is missing after examining I would recommend expanding acronyms in all figure captions for ease of reading.

We have rephrased this sentence to read:

"Examining sparse regression coefficients are informative for identifying important absorption bands, but interpretation can still be complicated by the compensation of interfering bands (Haaland and Thomas, 1988; Kvalheim et al., 2014)."

References

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, *Environmental Science & Technology*, 42, 4478–4485, doi:10.1021/es703009q, 2008.

Chhabra, P. S., Ng, N. L., Canagaratna, M. R., Corrigan, A. L., Russell, L. M., Worsnop, D. R., Flagan, R. C., and Seinfeld, J. H.: Elemental composition and oxidation of chamber organic aerosol, *Atmospheric Chemistry and Physics*, 11, 8827–8845, doi:10.5194/acp-11-8827-2011, 2011.

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: organic carbon, *Atmospheric Measurement Techniques*, 8, 1097–1109, doi:10.5194/amt-8-1097-2015, 2015a.

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance mea-

surements from infrared spectra: elemental carbon, *Atmospheric Measurement Techniques*, 8, 4013–4023, doi:10.5194/amt-8-4013-2015, 2015b.

Haaland, D. M. and Thomas, E. V.: Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, *Analytical Chemistry*, 60, 1193–1202, doi:10.1021/ac00162a020, 1988.

Kvalheim, O. M., Arneberg, R., Bleie, O., Rajalahti, T., Smilde, A. K., and Westerhuis, J. A.: Variable importance in latent variable regression models, *Journal of Chemometrics*, 28, 615–622, doi:10.1002/cem.2626, 2014.

Reggente, M., Dillner, A. M., and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: extending the predictions to different years and different sites, *Atmospheric Measurement Techniques*, 9, 441–454, doi:10.5194/amt-9-441-2016, 2016.

Ruggeri, G., Bernhard, F. A., Henderson, B. H., and Takahama, S.: Model-measurement comparison of functional group abundance in α-pinene and 1,3,5-trimethylbenzene secondary organic aerosol formation, *Atmospheric Chemistry and Physics Discussions*, 2016, 1–34, doi:10.5194/acp-2016-46, 2016.

Ruthenburg, T. C., Perlin, P. C., Liu, V., McDade, C. E., and Dillner, A. M.: Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the IMPROVE network, *Atmospheric Environment*, 86, 47–57, doi:10.1016/j.atmosenv.2013.12.034, 2014.

Takahama, S. and Dillner, A. M.: Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy, *Journal of Chemometrics*, 29, 659–668, doi:10.1002/cem.2761, 2015.