[27] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[28] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[29] Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

[30] Jaewook Lee and Daewon Lee. Dynamic characterization of cluster structures for robust and inductive support vector clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1869–1874, 2006.

[31] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[32] C Radhakrishna Rao. The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–358, 1964.

[33] H Farooqi, AM Sosa, PD Sottile, DJ Albers, and BJ Smith. Experimentally induced ventilator dyssynchrony increases injury during prolonged ventilation of endotoxin-injured mice. In *C72. HOUSE OF ARDS... AND MECHANICAL VENTILATORY SUPPORT*, pages A5780–A5780. American Thoracic Society, 2023.

[34] José M Amigó, Karsten Keller, and Valentina A Unakafova. Ordinal symbolic analysis and its application to biomedical recordings. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2034):20140091, 2015.

[35] Douglas Lind and Brian Marcus. *An introduction to symbolic dynamics and coding*. 2nd edition, 2021.

[36] Yoshito Hirata and José M Amigó. A review of symbolic dynamics and symbolic reconstruction of dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(5), 2023.

Appendix A. Supporting Figures

Appendix A.1. Intracluster normal and eFL in p111, label2

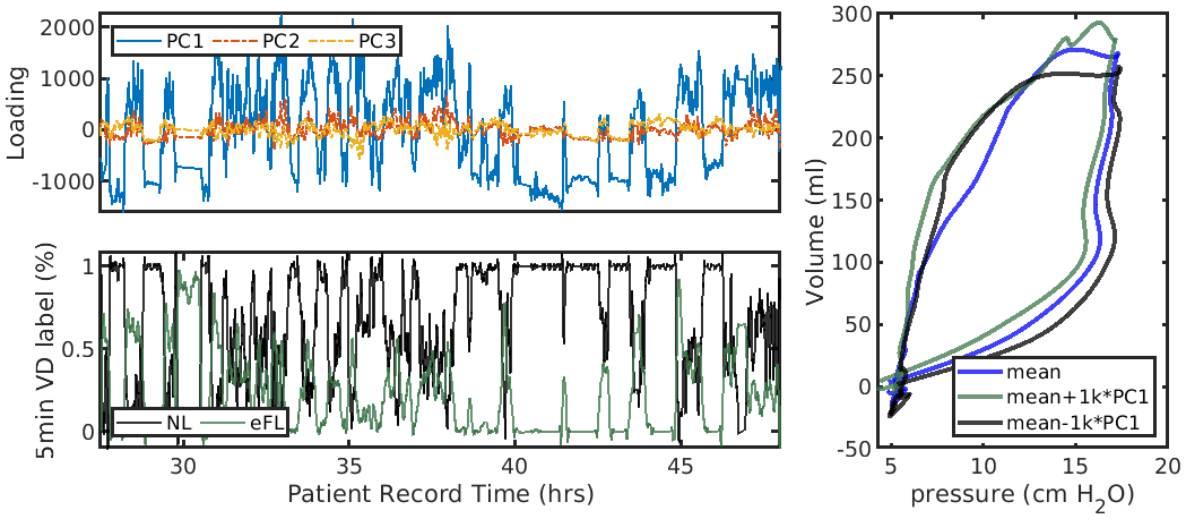

Figure A.7: The sign of PC1 loading roughly divides the VD classes in p111, label2. A threshold for the PC1 loading at zero roughly separates NL and eFL labels by 34%/65% and 85%/14%, respectively, with NL labels strongly associated with negative loadings. The optimal threshold (∼0.05) offers only subtle improvement. The right panel illustrates low fidelity changes in the cluster median pV loop (blue) when modified by these negative (black, more associated with NL) and positive (green, eFL) loadings. Note that this involves comprising 10-second properties (representing typically ∼3 breaths) to breathwise labels, and some representation errors thus arise from summarizing binary VD labels distributionally over all breaths intersecting a 10-second analysis window.

Appendix A.2. Outlier individual cluster characterizations in the cohort segmentation

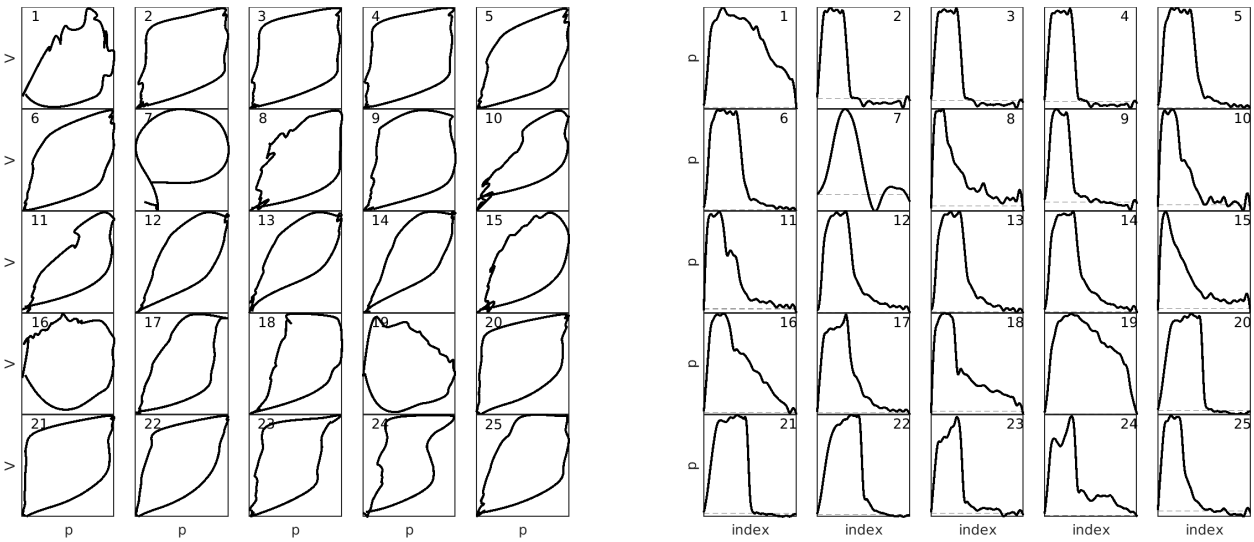

Figure A.8: Outlier pV (left) and pressure waveform (right) characterizations from two-stage cohort LVS phenotyping, shown in normalized form. Group categorizations associated with the largest 25 (of 27) outliers in Figure 6 which are distinct from the main identified groups. PEEP is approximated in noramlized pressure waveforms and indicated by dashed lines. Some outliers appear to be artifactual (from the data or estimation under stationarity). Others may be unique characterizations corresponding to extreme cases of VD, effects of patient posture, or heterogeneous breaths occurring under uncommonly used ventilator modes (e.g., spontaneous breathing present in 3.4% of breaths)

549 *Appendix A.3. Qualitiative equivalence of labels via tSNE & UMAP*

550 Methodological choices may bias the segmentation process of LVS descriptors. The feature dimensional
551 reduction method used prior to DBSCAN labeling is strongly influential on the labeling process. Cluster
552 labels are qualitatively the same in nearly all cases for under application of tSNE and UMAP (Figs. A.9
553 and A.10). However, extracted characterizations for populous groupings may differ due to the geometries
554 of embedded points. Characterization of tSNE-oriented labels appear to be more representative of realized
555 breaths: tSNE projection of features tend to be more convex, which results in mean and median points lying
556 closer to realized data. [[What i'm trying to say here: UMAP coordinates can be more asymmetric and less
ball-like with tentacles, and loss of convexity means the 'center' can lie farther from the actual features.]]

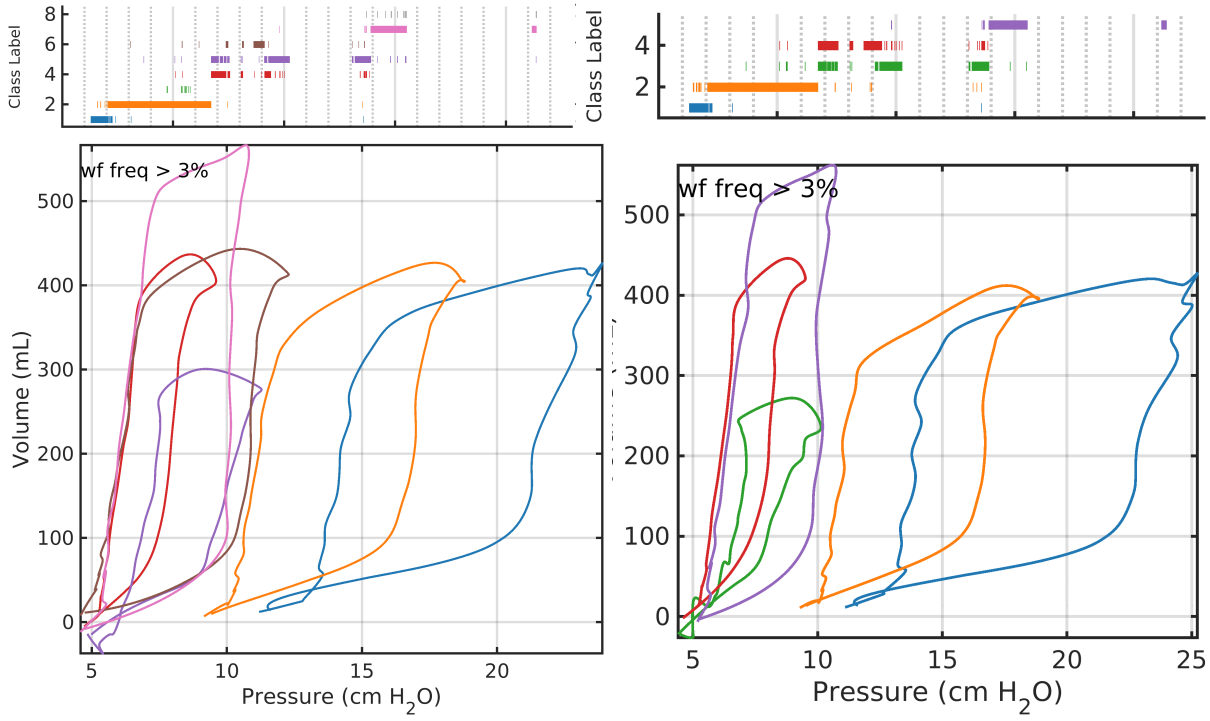

Figure A.9: Patient 101 clustering using tSNE (left) and UMAP (right) feature reduction stages. Identified phenotypes show qualitatively similar evolution although the tSNE-based characterization are more representative due poor representation of non-convex UMAP groupings by the component-wise median.

557

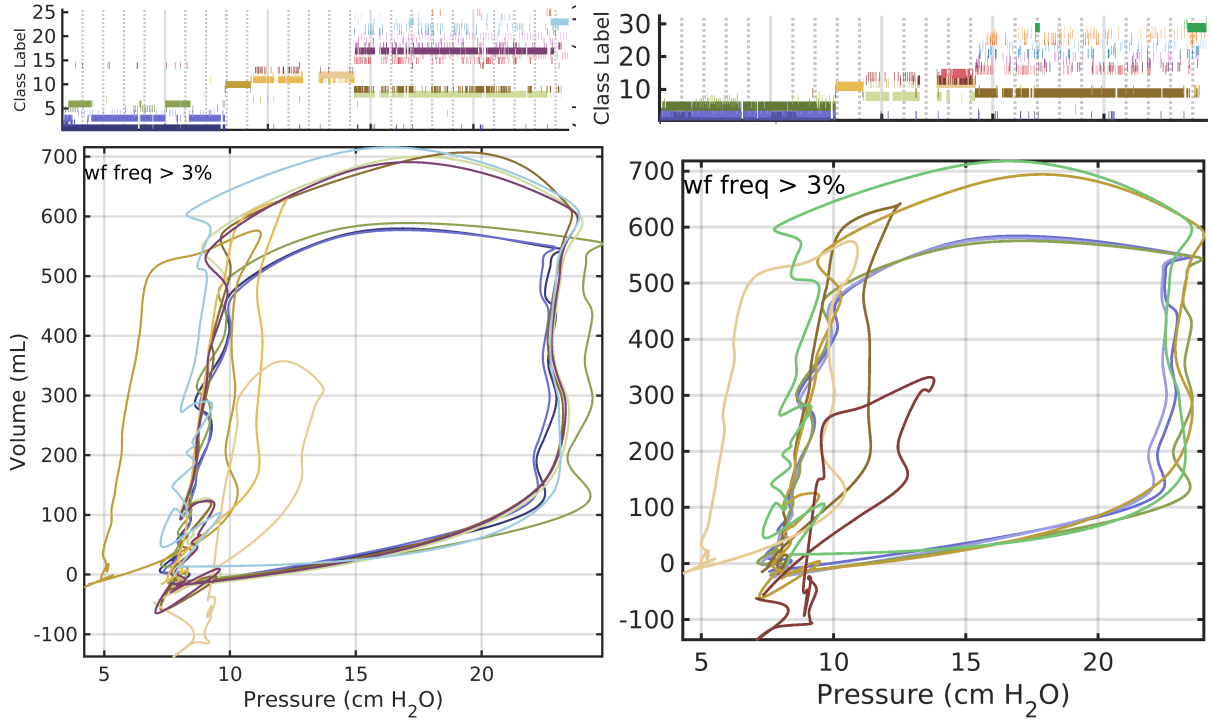

Figure A.10: Patient 124 clustering, as above

Appendix B. Sample size vs. Sample description

Broadly, waveform digitization transforms high-frequency temporal sampling of state processes into a lower-frequency, distributionally-descriptive form. This reduces the effective size of the problem while making it more dense. For a classification problem involving T samples of M-dimensional observations stored in an array $D \in \mathbb{R}^{T\times M}$, methods involving kernel or covariance processes require then calculating a matrix of dimension $M \times (T \times T) \times M$ in observation space or $T \times (M \times M) \times T$ in sample space. Decreasing the order of T and increasing that of M by a factor α benefits computational efficiency by replacing $D \in \mathbb{R}^{T\times M}$ with $\widetilde{D} \in \mathbb{R}^{(T/\alpha)\times(\alpha M)}$. Specifically, calculating the observation covariance from $\widetilde{D}$ requires α^2 more storage but involves α^{-2} fewer calculations over the samples:[wording bad:] $(\alpha M)\times(\alpha M)$ is calculated via $\alpha^{-1}T \times \alpha^{-1}T$ rather than $T \times T$. Similarly, the summary sample space covariance of size $(T/\alpha) \times (T/\alpha)$ may be more dense than one built from un-summarized samples in $T \times T$, but it may be machine representable for larger values of T. Computational effects are important as $T >> M$ in most practical applications, and additional statistical benefits arise from increasing the size of M.