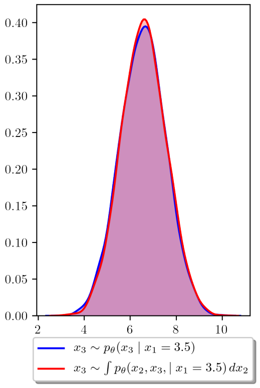
Motivation of the method Following Reviewer #3, we clarify the motivation behind NC. We emphasize our goal of learning features from unsupervised data, useful for downstream tasks. This is the main point validated throughout our experiments. We show that our paradigm of learning many conditional distributions of the data allows the extraction of these unsupervised features from incomplete data, as well as arbitrary data imputations (inpaintings).

General changes to the manuscript Following Reviewer's #1 suggestion, we included in the Appendix our experimental protocols, architectures, and optimization parameter grids for all methods. We also added error bars to Tables 1 and 3 from the manuscript. Following Reviewer's #3 suggestion, we tone down our claims about unsupervised learning.

Regarding Sections 4.2 and 4.3 We apologize for the confusion brought by using the Donsker-Varadhan lower bound as an objective for the discriminator. As pointed out by Reviewer #1, we acknowledge that the statement from lines 188-190 is misleading. Our intent was to reason about optimal discriminators. Given an optimal discriminator, the optimal NC minimizes the Jensen-Shannon divergence between $p_{\theta^*}(x_r \mid x_a, a, r)$ and $p(x_r \mid x_a, a, r)$, where p_θ and p represent the model and data distributions, respectively. Consequently, at optimality we have that $D_{KL}(p||p_{\theta^*}) = 0$, and thus the negative log-likelihood is equal to $H(X_R|X_A)$. Then, the more information X_A holds about X_R, the lower the negative log-likelihood. Following Reviewer's #1 and #3 remarks, we replace the Donsker-Varadhan lower bound by one in terms of the Jensen-Shannon divergence, merging Sections 4.2 and 4.3, removing the misleading statement from lines 188-190, making clear that our reasoning follows for optimal discriminators and NC's, and making the variables on which the different distributions depend explicit. We thank the reviewers for their careful reading.

Regarding conditional distributions consistency We thank Reviewer #1 for bringing this subtle point to our attention. We now provide a proof about the consistency of conditional and marginalized densitities in our Appendix. The proof sketch goes as follows: if we assume that the data has support on a compact set Ω, and that the NC is trained to optimality, then, writing λ for the Lebesgue measure on Ω, we can show that $D_{JS}(p_\theta(x_S \mid x_A)|| \int p_\theta(x_S, x_{R-S} \mid x_A)\lambda(dx_{R-S}))$ is small by leveraging the triangular inequality of the distance on probability measure on Ω defined by the square root of the Jensen-Shannon divergence. We then use Jensen's inequality with uniform weights $\frac{1}{\lambda(supp(X_{R-S}))}$ to bound the distance between the model and data marginalized distributions by the integral of the distance between the model and data conditional distributions. We leave the theoretical analysis for the case for non-compact supports to future work. Moreover, we illustrate this consistency between conditional and marginalized densities empirically, here in the Figure on the right.



Better empirical comparison against VAEAC As suggested, we improve our empirical comparison against VAEAC and update the manuscript with the results shown here in Table 1 (Left, semi-supervised learning results on SVHN), and Table 1 (Middle, missing data imputation on three UCI datasets).

Improved performances on the missing data imputation task As requested by Reviewer #3, Table 1 (Middle) shows substantially improved missing data imputations results for our model. This results were obtained after fixing a bug in our code.

Performance degradation depending on train/test masks mismatch We follow Reviewer's #1 suggestion to analyze the performance of NC as a function of the mismatch between train and test masks. To this end, we report the RMSE between X_R and $\hat{X}_R$ on the UCI/Letter dataset. We consider four different scenarios: (i) masks observed during training applied on train data, (ii) masks observed during training applied on test data, (iii) masks unobserved during training applied on train data, and (iv) masks unobserved during training applied on test data. Results in Table 1 (Right).

Algorithm	Test error (%)
VAEAC	57.89 ± 1.01
NC	**17.2 ± 0.59**

Algorithm	Spam	Letter	Credit
GAIN	.0513 ± .002	.1198 ± .005	.1858 ± .001
VAEAC	.0552 ± .002	.1115 ± .001	.1523 ± .002
NC	**.0486 ± .001**	**.0851 ± .002**	**.1276 ± .002**

Masks/Data	Train	Test
Train	.0891 ± .001	.0896 ± .001
Test	.0897 ± .001	.0901 ± .001

Table 1: **(Left)** Semi-supervised learning on SVHN using 1000 labels. **(Middle)** RMSE for missing data imputation on UCI datasets. **(Right)** RMSE on UCI/Letter using train/test data/masks.