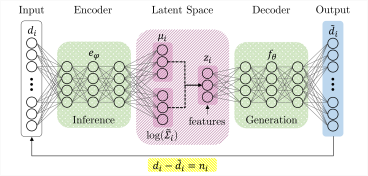
Probabilistic Variational Autoencoders

A Bayesian perspective



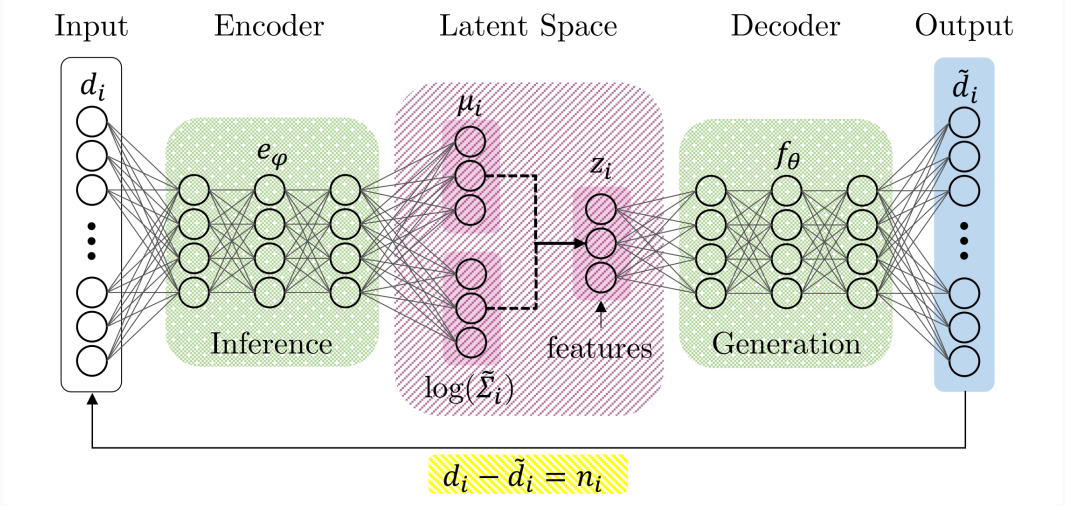
Philipp Frank

February 7, 2022

Berlin Machine Learning Group, Germany

www.ph-frank.de

Variational Autoencoder (VAE)

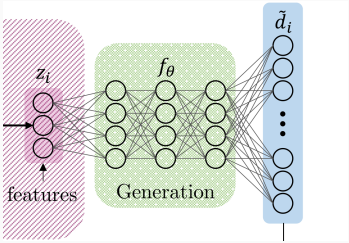


Probabilistic generative model

VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$

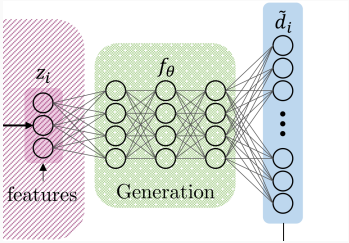
$\mathcal{P}(z) \sim z_i$



VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$
- Generative model $\tilde{d} = f_\theta(z)$

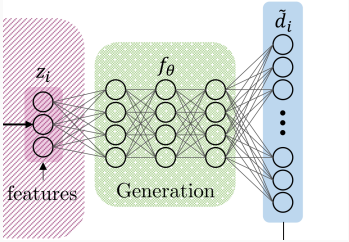
$\mathcal{P}(z) \sim z_i$



VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$
- Generative model $\tilde{d} = f_\theta(z)$
- Probabilistic model $d = \tilde{d} + n; \quad n \sim \mathcal{P}(n)$

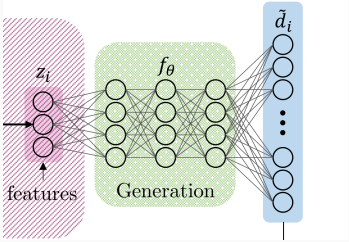
$\mathcal{P}(z) \sim z_i$



VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$
- Generative model $\tilde{d} = f_\theta(z)$
- Probabilistic model $d = \tilde{d} + n; \quad n \sim \mathcal{P}(n)$
- E.g. $\mathcal{P}(n) = \mathcal{N}(n|0, \mathbb{1})$

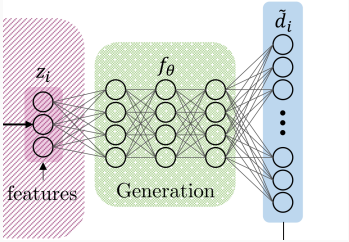
$\mathcal{P}(z) \sim z_i$



VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$
- Generative model $\tilde{d} = f_\theta(z)$
- Probabilistic model $d = \tilde{d} + n; \quad n \sim \mathcal{P}(n)$
- E.g. $\mathcal{P}(n) = \mathcal{N}(n|0, N)$

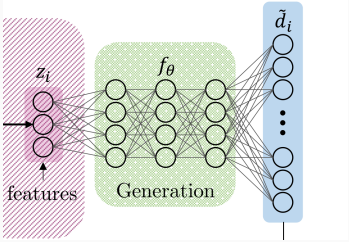
$\mathcal{P}(z) \sim z_i$



VAE - Generative model (Decoder)

- Latent features from common distribution $\mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$
- Generative model $\tilde{d} = f_\theta(z)$
- Probabilistic model $d = \tilde{d} + n; \quad n \sim \mathcal{P}(n)$
- E.g. $\mathcal{P}(n) = \mathcal{N}(n|0, N)$
- Adaptive covariance $N \sim \mathcal{P}(N)$

$\mathcal{P}(z) \sim z_i$



Inference

VAE - Inference

- Likelihood $\mathcal{P}\left(d_i|\theta, z_i\right) = \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)$
- Prior $\mathcal{P}\left(\theta, z_i\right) = \mathcal{P}(\theta)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)$

VAE - Inference

- Likelihood $\mathcal{P}\left(d_i|\theta, z_i\right) = \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)$
- Prior $\mathcal{P}\left(\theta, z_i\right) = \mathcal{P}(\theta)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)$
- Dataset $\mathcal{D} \equiv \{d_i\}_{i\in\{1,\ldots,M\}}$; Latent variables $\mathcal{Z} \equiv \{z_i\}_{i\in\{1,\ldots,M\}}$

VAE - Inference

- Likelihood $\mathcal{P}\left(d_i|\theta, z_i\right) = \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)$
- Prior $\mathcal{P}\left(\theta, z_i\right) = \mathcal{P}(\theta)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)$
- Dataset $\mathcal{D} \equiv \{d_i\}_{i\in\{1,\ldots,M\}}$; Latent variables $\mathcal{Z} \equiv \{z_i\}_{i\in\{1,\ldots,M\}}$

Product Rule of Probabilities aka Bayes' theorem

$$\mathcal{P}\left(\theta, \mathcal{Z}|\mathcal{D}\right) = \frac{\left[\prod_{i=1}^{M} \mathcal{P}\left(d_i|\theta, z_i\right)\ \mathcal{P}\left(z_i\right)\right]\mathcal{P}(\theta)}{\mathcal{P}\left(\mathcal{D}\right)}\ .$$

VAE - Inference

- Likelihood $\mathcal{P}\left(d_i|\theta, z_i\right) = \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)$
- Prior $\mathcal{P}\left(\theta, z_i\right) = \mathcal{P}(\theta)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)$
- Dataset $\mathcal{D} \equiv \{d_i\}_{i\in\{1,\ldots,M\}}$; Latent variables $\mathcal{Z} \equiv \{z_i\}_{i\in\{1,\ldots,M\}}$

Product Rule of Probabilities aka Bayes' theorem

$$\mathcal{P}\left(\theta, \mathcal{Z}|\mathcal{D}\right) = \frac{\left[\prod_{i=1}^{M} \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)\right] \mathcal{P}(\theta)}{\mathcal{P}\left(\mathcal{D}\right)}\ .$$

VAE - Inference

- Likelihood $\mathcal{P}\left(d_i|\theta, z_i, N\right) = \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)$
- Prior $\mathcal{P}\left(\theta, z_i, N\right) = \mathcal{P}(\theta)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)\ \mathcal{P}(N)$
- Dataset $\mathcal{D} \equiv \{d_i\}_{i\in\{1,\dots,M\}}$; Latent variables $\mathcal{Z} \equiv \{z_i\}_{i\in\{1,\dots,M\}}$

Product Rule of Probabilities aka Bayes' theorem

$$\mathcal{P}\left(\theta, \mathcal{Z}, N|\mathcal{D}\right) = \frac{\left[\prod_{i=1}^{M} \mathcal{N}\left(d_i|f_\theta\left(z_i\right), N\right)\ \mathcal{N}\left(z_i|0, \mathbb{1}\right)\right] \mathcal{P}(\theta)\ \mathcal{P}(N)}{\mathcal{P}\left(\mathcal{D}\right)}\ .$$

Variational Approximation

- Posterior distribution $\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})$ intractable

Variational Approximation

- Posterior distribution $\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})$ intractable
- Approximation with tractable distribution $\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})$

Variational Approximation

- Posterior distribution $\mathcal{P}\left(\theta, \mathcal{Z}|\mathcal{D}\right)$ intractable
- Approximation with tractable distribution $\mathcal{Q}\left(\theta, \mathcal{Z}|\mathcal{D}\right)$

Kullback-Leibler Divergence

$$\mathcal{KL}\left[\mathcal{Q};\ \mathcal{P}\right] \equiv -\int \log\left(\frac{\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})}{\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})}\right) \mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})\ \mathrm{d}\theta\ \mathrm{d}\mathcal{Z}\ .$$

Variational Approximation

- Posterior distribution $\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})$ intractable
- Approximation with tractable distribution $\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})$

Kullback-Leibler Divergence

$$\mathcal{KL}[\mathcal{Q};\ \mathcal{P}] \equiv -\int \log\left(\frac{\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})}{\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})}\right) \mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})\ \mathrm{d}\theta\ \mathrm{d}\mathcal{Z}\ .$$

- Parameterize approximation: $\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D}) = \delta\left(\theta - \bar{\theta}\right) \prod_{i=1}^{M} \mathcal{Q}_\phi\left(z_i|d_i\right)$

Variational Approximation

- Posterior distribution $\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})$ intractable
- Approximation with tractable distribution $\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})$

Kullback-Leibler Divergence

$$\mathcal{KL}[\mathcal{Q};\ \mathcal{P}] \equiv -\int \log\left(\frac{\mathcal{P}(\theta, \mathcal{Z}|\mathcal{D})}{\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})}\right) \mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D})\ \mathrm{d}\theta\ \mathrm{d}\mathcal{Z}\ .$$

- Parameterize approximation: $\mathcal{Q}(\theta, \mathcal{Z}|\mathcal{D}) = \delta\left(\theta - \bar{\theta}\right) \prod_{i=1}^{M} \mathcal{Q}_\phi\left(z_i|d_i\right)$
- Optimize the $\mathcal{KL}$ for $\bar{\theta}$ and ϕ

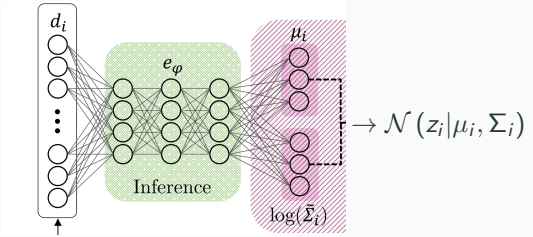
Inference Network (Encoder)

- Parameterize as a normal distribution
  $\mathcal{Q}_\phi(z_i|d_i) \equiv \mathcal{N}(z_i|\mu_i, \Sigma_i)$

$\rightarrow \mathcal{N}(z_i|\mu_i, \Sigma_i)$

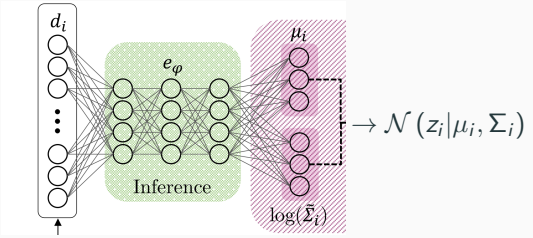
Inference Network (Encoder)

- Parameterize as a normal distribution $\mathcal{Q}_\phi(z_i|d_i) \equiv \mathcal{N}(z_i|\mu_i, \Sigma_i)$
- And inference network $\begin{pmatrix} \mu_i \\ \Sigma_i \end{pmatrix} = e_\phi(d_i)$



Inference Network (Encoder)

- Parameterize as a normal distribution $\mathcal{Q}_\phi(z_i|d_i) \equiv \mathcal{N}(z_i|\mu_i, \Sigma_i)$
- And inference network $\begin{pmatrix}\mu_i \\ \Sigma_i\end{pmatrix} = e_\phi(d_i)$
- Σ often parameterized as a diagonal matrix



Extensions / Modifications

VAE - Extensions

- Exchange / Modify likelihood distribution

VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...

VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Alter approximation distribution $\mathcal{Q}$

VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Alter approximation distribution $\mathcal{Q}$
  - Use full covariance matrix

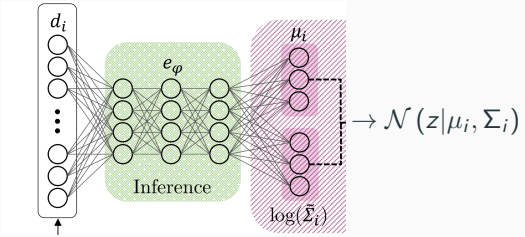
VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Alter approximation distribution $\mathcal{Q}$
  - Use full covariance matrix
  - Estimate covariance from generative model using Fisher Information metric
  - ...

Fisher metric as covariance

Fisher Information metric

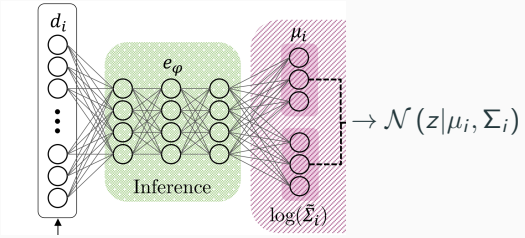
$$\mathcal{M}(\mu) \equiv -\left\langle \frac{\partial^2 \log\left(\mathcal{P}\right)\left(d|z,\theta\right)}{\partial z\ \partial z} \right\rangle_{\mathcal{P}(d|z,\theta)}\Bigg|_{z=\mu} + \mathbb{1}$$



Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

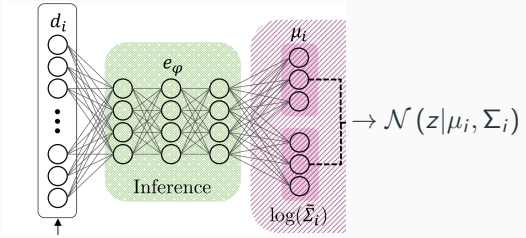


Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$

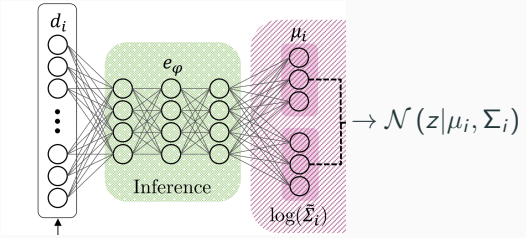


Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$
- Fisher-Metric is an approximation to post. covariance (including off-diagonal)

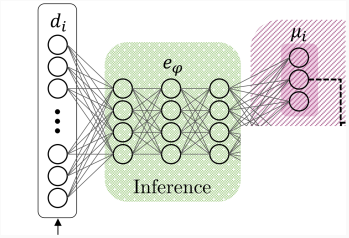


Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$
- Fisher-Metric is an approximation to post. covariance (including off-diagonal)

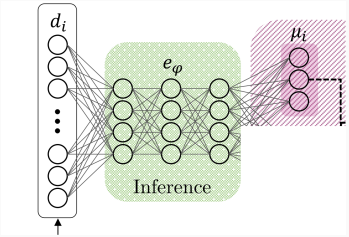


Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$
- Fisher-Metric is an approximation to post. covariance (including off-diagonal)
- Coupled second moments of $\mathcal{P}$ and $\mathcal{Q}$

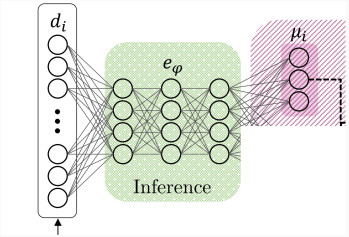


Fisher metric as covariance

Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$
- Fisher-Metric is an approximation to post. covariance (including off-diagonal)
- Coupled second moments of $\mathcal{P}$ and $\mathcal{Q}$
- Increased inference capabilities of $\mathcal{Q}$

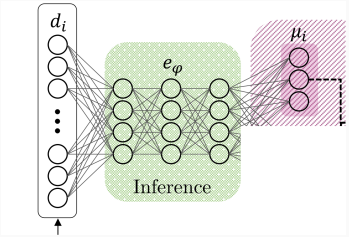


Fisher metric as covariance

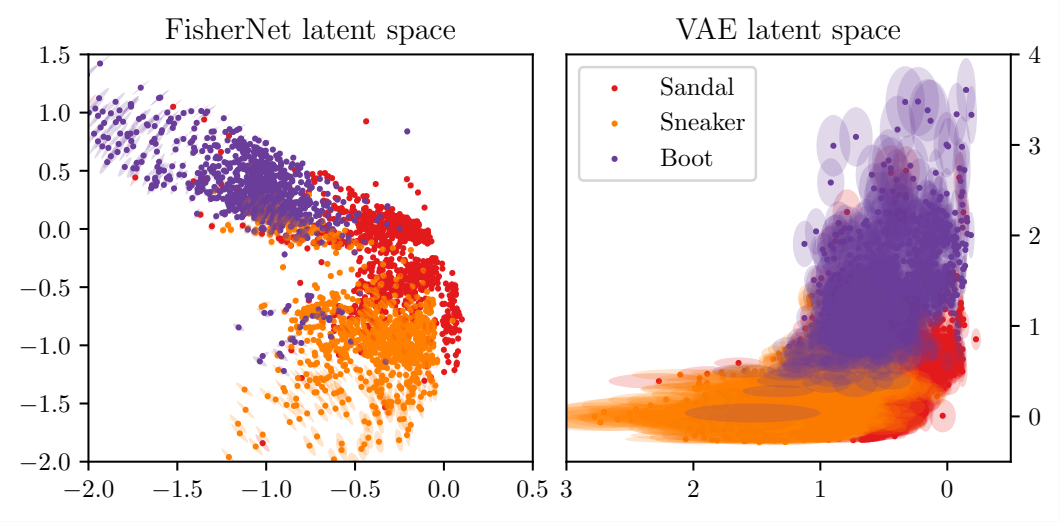
Fisher Information metric

$$\mathcal{M}(\mu) \equiv \left[\left(\frac{\partial f_\theta}{\partial z} \right)^\dagger N^{-1} \left(\frac{\partial f_\theta}{\partial z} \right) \right]_{z=\mu} + \mathbb{1}$$

- Kramer-Rao bound
  $\mathcal{M}^{-1}(\mu) \leq \left\langle (z-\mu)(z-\mu)^\dagger \right\rangle_{\mathcal{P}(z|d)}$
- Fisher-Metric is an approximation to post. covariance (including off-diagonal)
- Coupled second moments of $\mathcal{P}$ and $\mathcal{Q}$
- Increased inference capabilities of $\mathcal{Q}$
- Sampling is less trivial



Comparison - Fashion MNIST



VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Change approximation distribution $\mathcal{Q}$
  - Non-Gaussian approximation
  - Use full covariance matrix
  - Estimate covariance from generative model using Fisher Information metric
  - ...

VAE - Extensions

- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Change approximation distribution $\mathcal{Q}$
  - Non-Gaussian approximation
  - Use full covariance matrix
  - Estimate covariance from generative model using Fisher Information metric
  - ...
- Post-process encoded distribution $\mathcal{Q}$ in latent space

VAE - Extensions

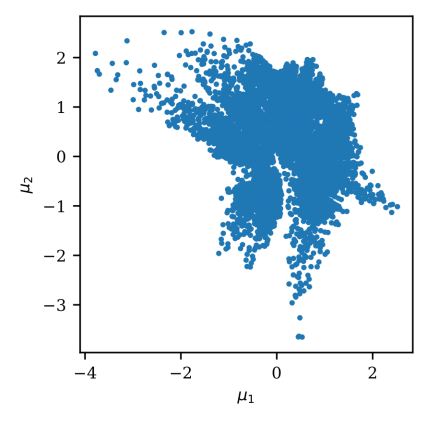
- Exchange / Modify likelihood distribution
  - Noise covariance N becomes a variable
  - ...
- Change approximation distribution $\mathcal{Q}$
  - Non-Gaussian approximation
  - Use full covariance matrix
  - Estimate covariance from generative model using Fisher Information metric
  - ...
- Post-process encoded distribution $\mathcal{Q}$ in latent space
  - Transform latent space distribution to a standard normal distribution for sample generation
  - ...

Transform latent space

- Marginal approximate prior distribution $\mathcal{Q}(z) = \int \mathcal{Q}(z|d)\ P(d)\ \mathrm{d}d$

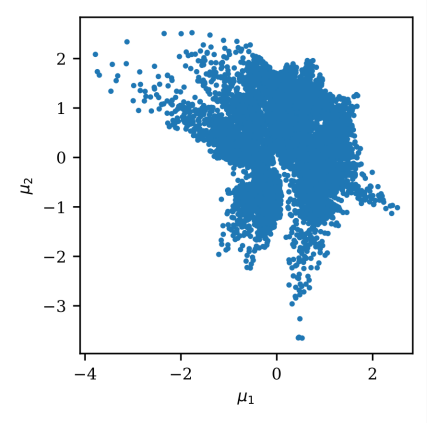
Transform latent space

- Marginal approximate prior distribution $\mathcal{Q}(z) = \int \mathcal{Q}(z|d)\ P(d)\ \mathrm{d}d$



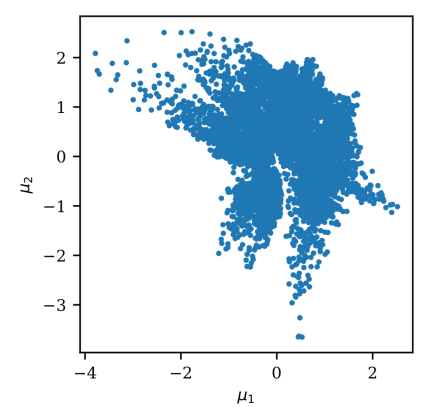
Transform latent space

- Marginal approximate prior distribution $\mathcal{Q}(z) = \int \mathcal{Q}(z|d)\ P(d)\ \mathrm{d}d$
- Ideally $\mathcal{Q}(z) \simeq \mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$

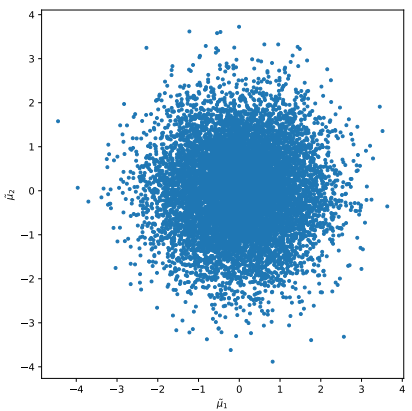


Transform latent space

- Marginal approximate prior distribution $\mathcal{Q}(z) = \int \mathcal{Q}(z|d)\ P(d)\ \mathrm{d}d$
- Ideally $\mathcal{Q}(z) \simeq \mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$

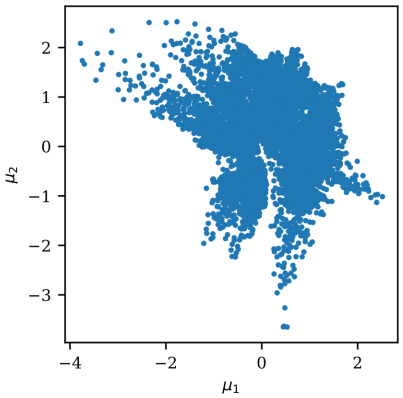


Transform using e.g. Normalizing flows or density estimation



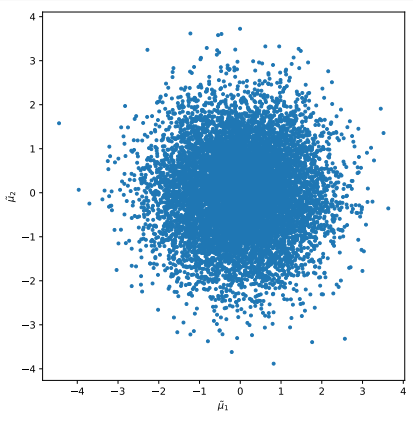
Transform latent space

- Marginal approximate prior distribution $\mathcal{Q}(z) = \int \mathcal{Q}(z|d)\ P(d)\ \mathrm{d}d$
- Ideally $\mathcal{Q}(z) \simeq \mathcal{P}(z) = \mathcal{N}(z|0, \mathbb{1})$



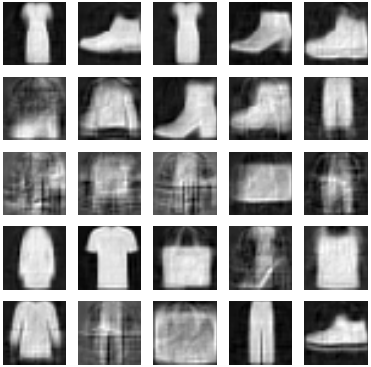
Transform using e.g. Normalizing flows or density estimation

Use inverse transformation for sample generation



Transform latent space

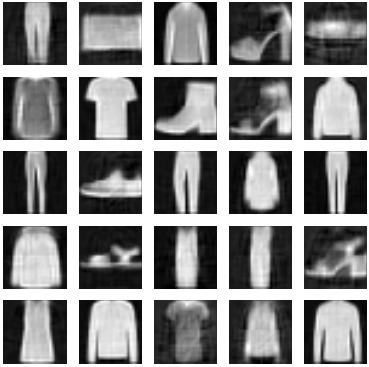
Drawn from latent space



FID: 110

(Training FID: 80)

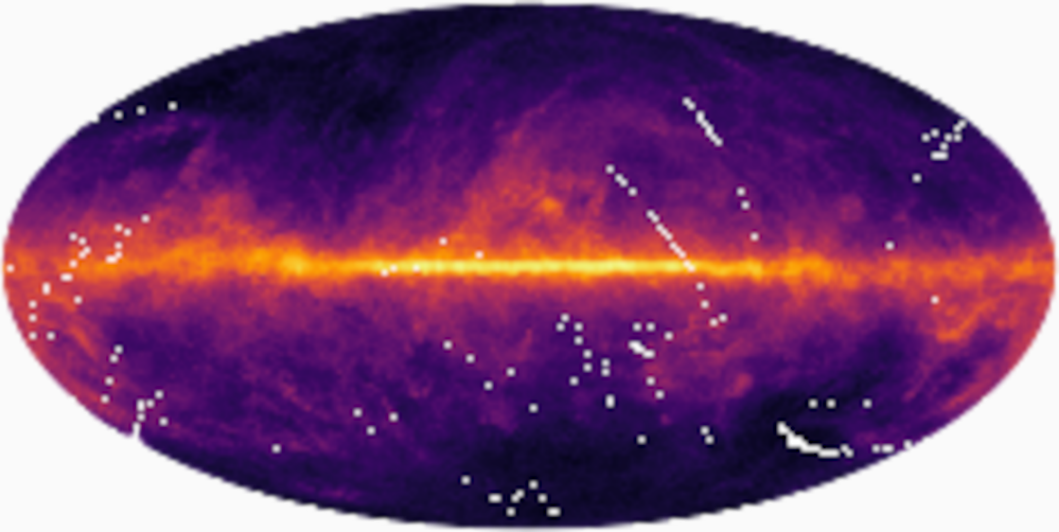
Drawn from transformed space



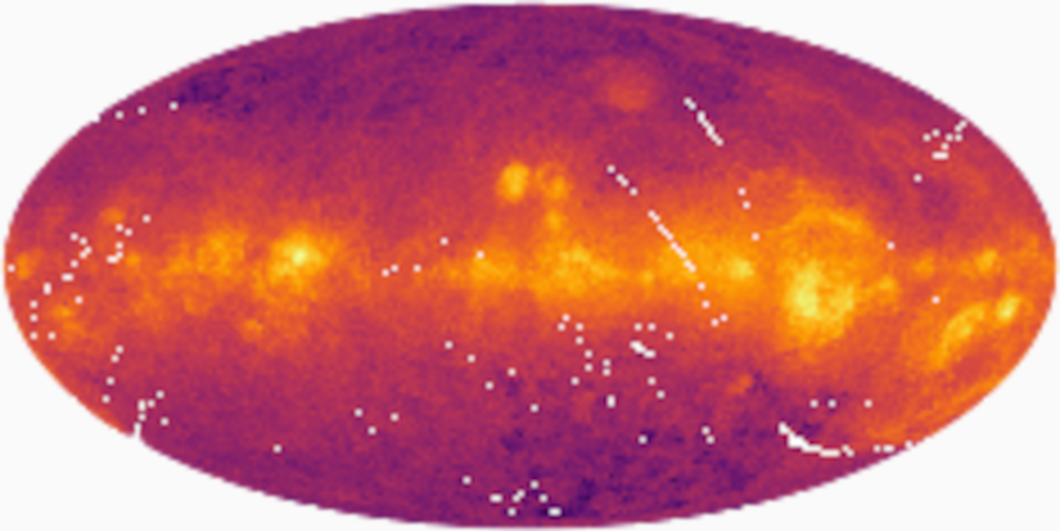
FID: 85

Decomposition of the Galactic multi-frequency sky

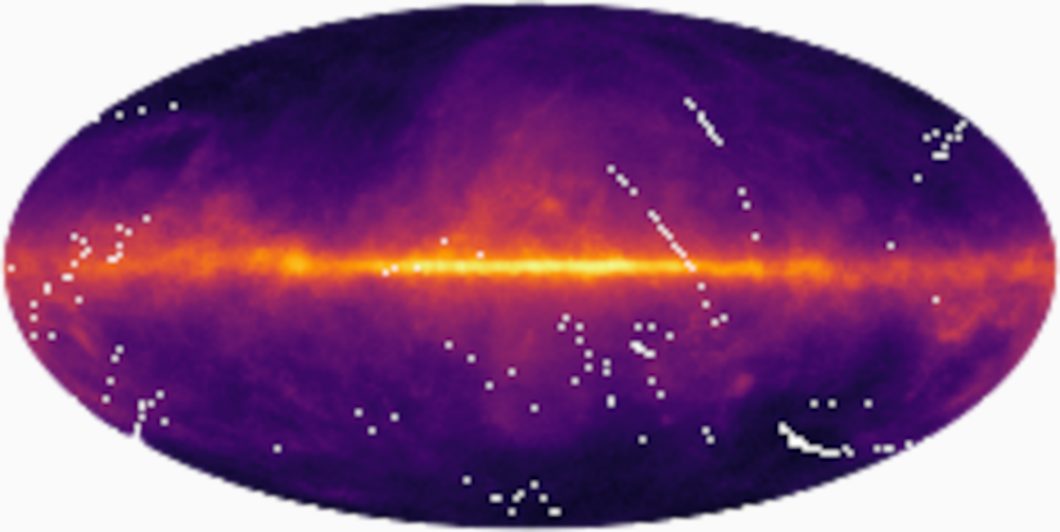
Planck Map (545 GHz)



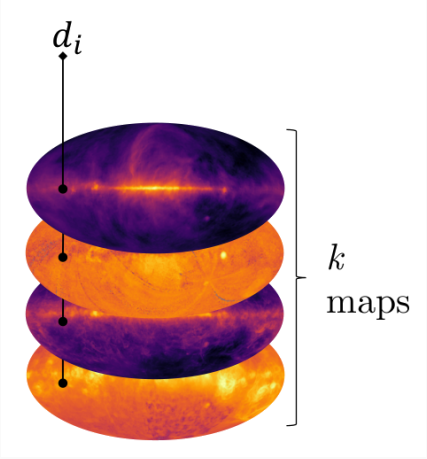
H-alpha (656.3 nm)



Fermi γ-ray map (1.7 GeV)

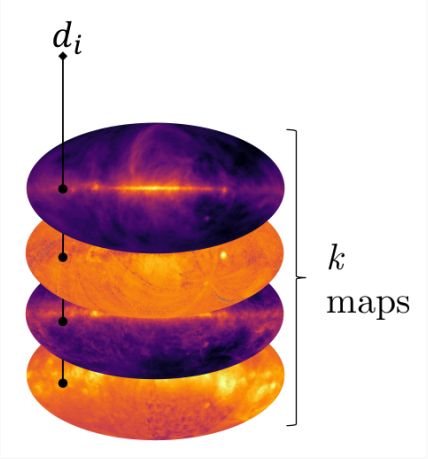


Decomposition of the Galactic multi-frequency sky - Data



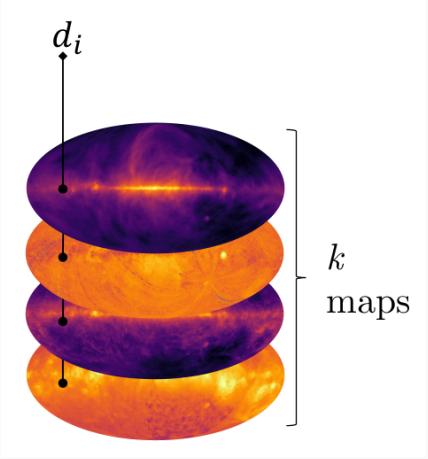
Decomposition of the Galactic multi-frequency sky - Data

- Pixel based approach (d_i: frequency brightness vector at ith-location on the sky)



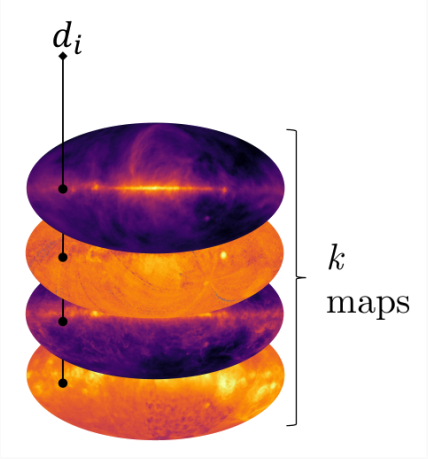
Decomposition of the Galactic multi-frequency sky - Data

- Pixel based approach (d_i: frequency brightness vector at ith-location on the sky)
- $k = 35$ all-sky maps



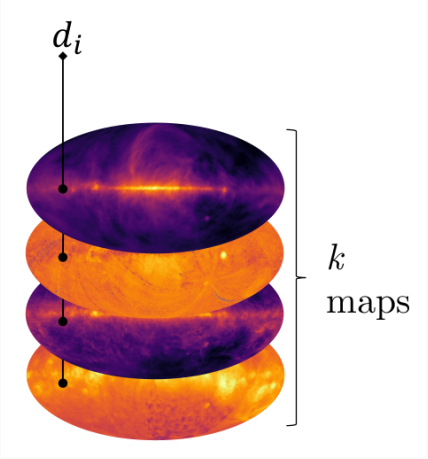
Decomposition of the Galactic multi-frequency sky - Data

- Pixel based approach (d_i: frequency brightness vector at ith-location on the sky)
- $k = 35$ all-sky maps
- Frequencies from Radio (MHz) to gamma ray (GeV)

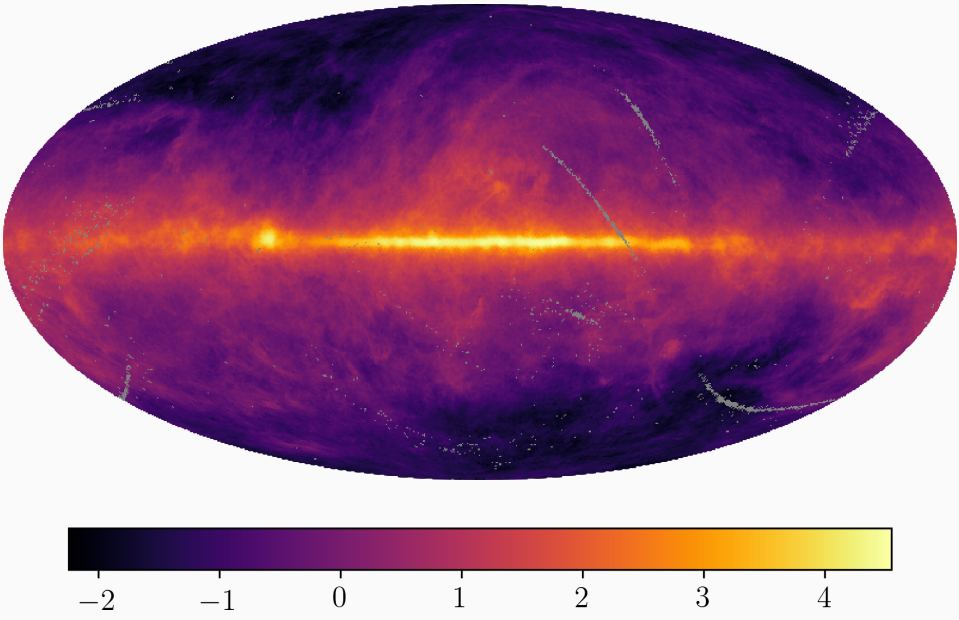


Decomposition of the Galactic multi-frequency sky - Data

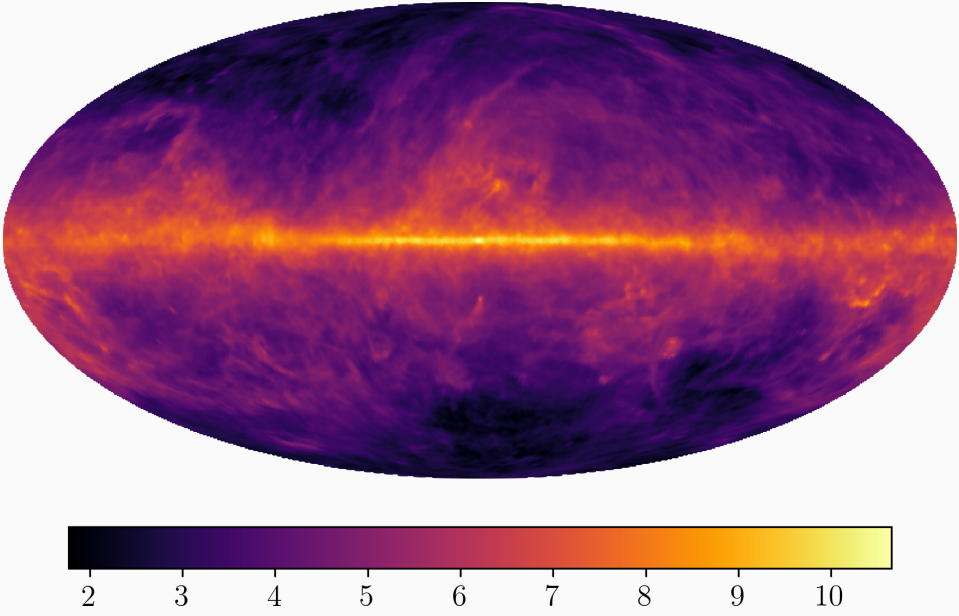
- Pixel based approach (d_i: frequency brightness vector at ith-location on the sky)
- $k = 35$ all-sky maps
- Frequencies from Radio (MHz) to gamma ray (GeV)
- 10 dimensional latent space



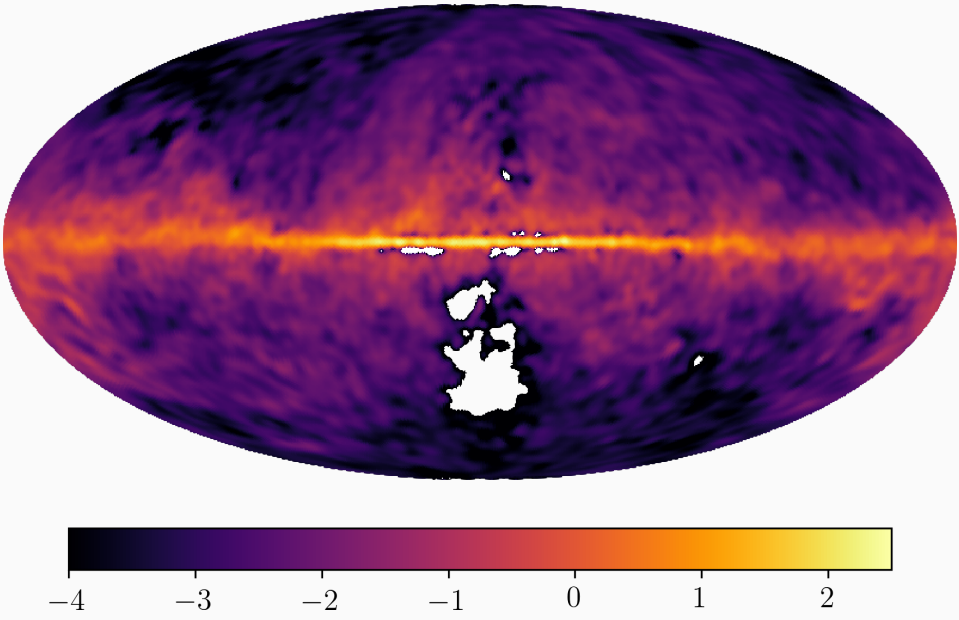
Decomposition of the Galactic multi-frequency sky - Feature A



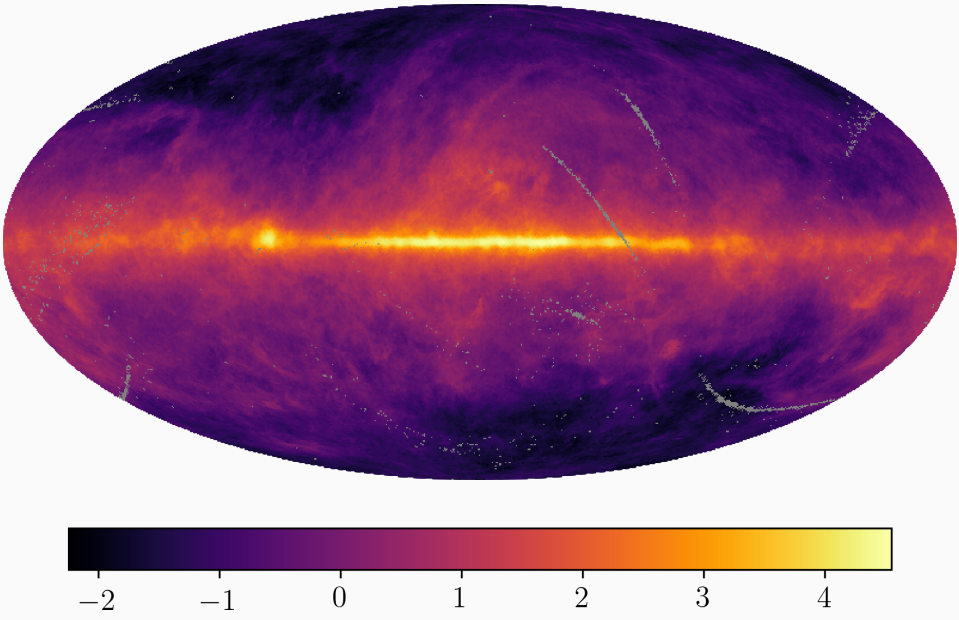
Decomposition of the Galactic multi-frequency sky - Dust



Decomposition of the Galactic multi-frequency sky - Hadronic Component

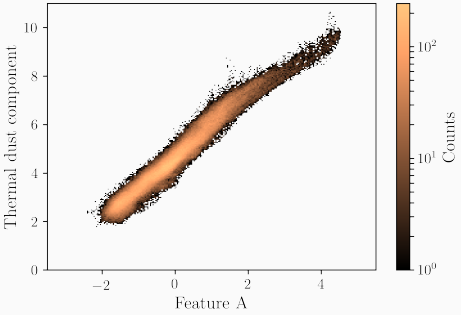


Decomposition of the Galactic multi-frequency sky - Feature A



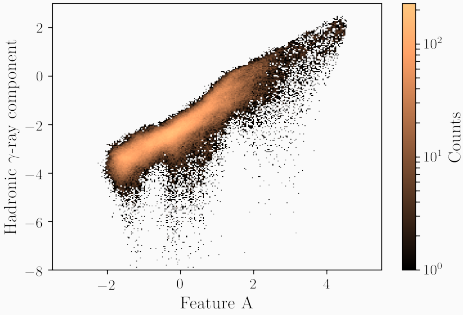
Decomposition of the Galactic multi-frequency sky - Correlations

Correlation Dust



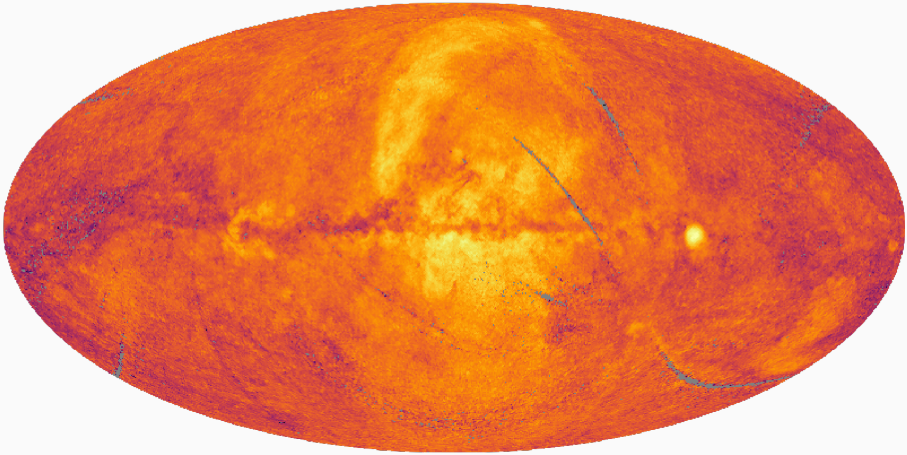
Mutual Information: $I(X;Y) = 1.72$

Correlation Hadronic Component



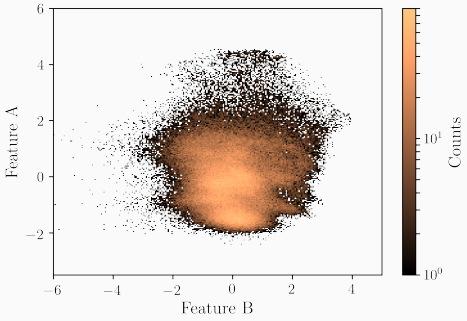
Mutual Information: $I(X;Y) = 1.07$

Decomposition of the Galactic multi-frequency sky - Feature B



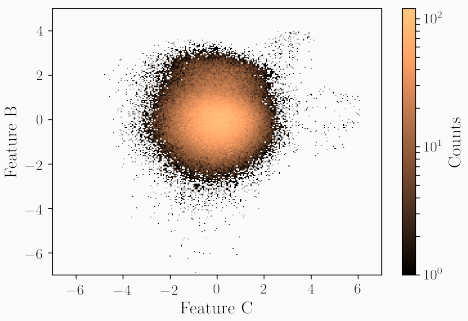
Decomposition of the Galactic multi-frequency sky - Correlations

Correlation Feature A



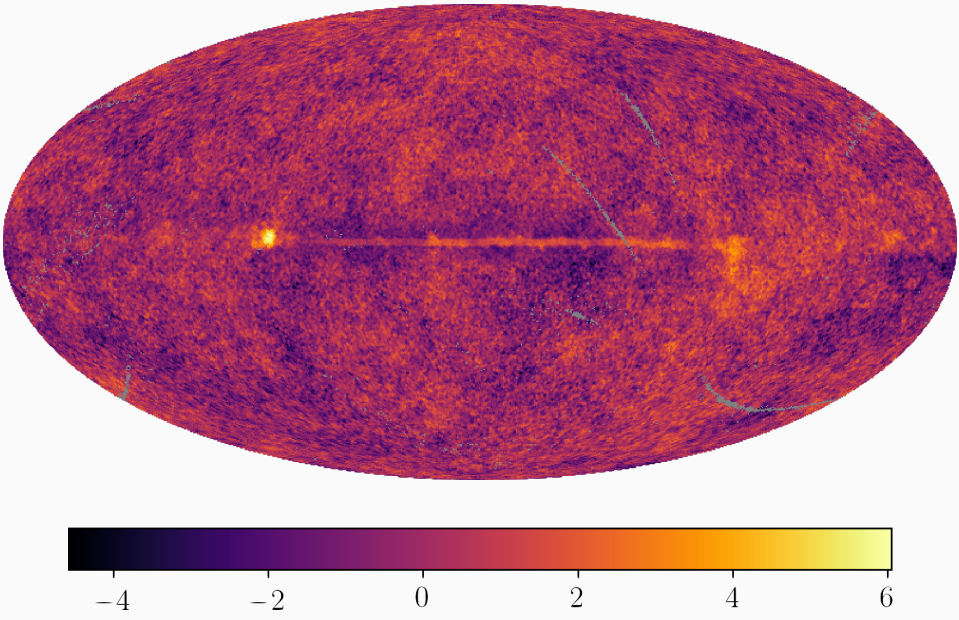
Mutual Information: $I(X;Y) = 0.33$

Correlation Feature C

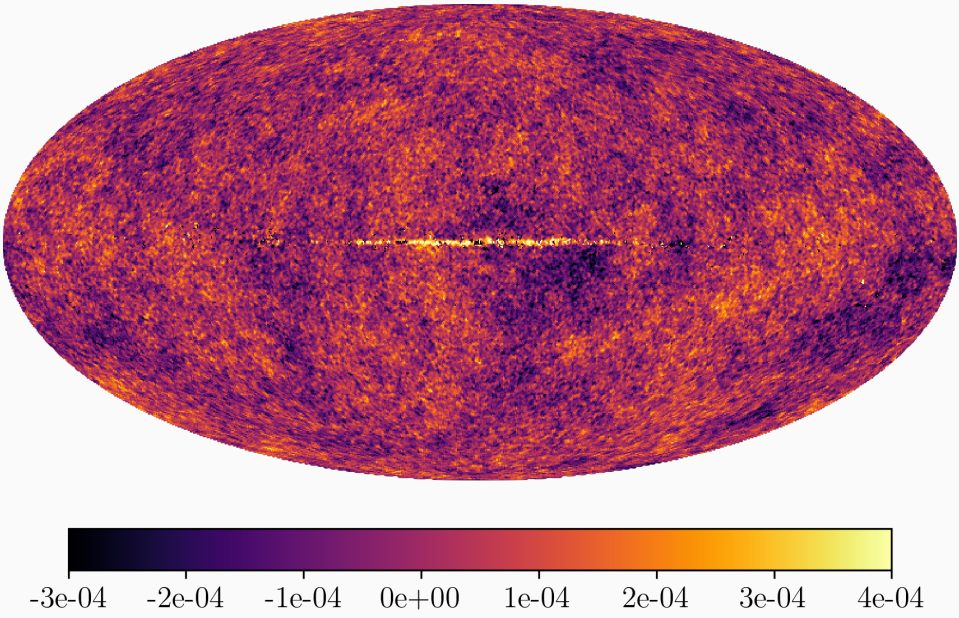


Mutual Information: $I(X;Y) = 0.20$

Decomposition of the Galactic multi-frequency sky - Feature C

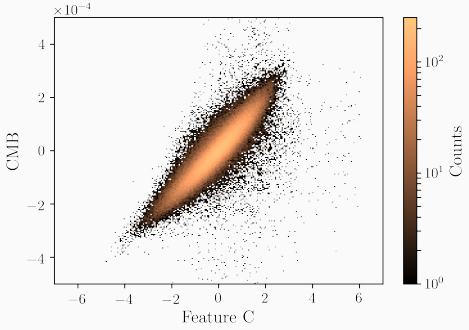


Decomposition of the Galactic multi-frequency sky - CMB



Decomposition of the Galactic multi-frequency sky - Correlations

Correlation CMB



Mutual Information: $I(X;Y) = 0.88$

Summary

- VAEs use probabilistic generative models

Summary

- VAEs use probabilistic generative models
- Associated inverse problem solved using variational inference

Summary

- VAEs use probabilistic generative models
- Associated inverse problem solved using variational inference
- Fisher information metric as covariance can improve inference capacity
- Posterior analysis of latent space using normalizing transformations can improve sampling quality

Summary

- VAEs use probabilistic generative models
- Associated inverse problem solved using variational inference
- Fisher information metric as covariance can improve inference capacity
- Posterior analysis of latent space using normalizing transformations can improve sampling quality
- Latent features of multi-frequency sky partially coincide with known physical components superimposed on the sky

References

- **Bayesian decomposition of the Galactic multi-frequency sky using probabilistic autoencoders**; Milosevic, S., Frank, P., Leike, R., Müller, A., Enßlin, T.A.; Astronomy & Astrophysics 2021, 650, A100. doi.org/10.1051/0004-6361/202039435
- **Probabilistic Autoencoder using Fisher Information**; Zacherl, J., Frank, P., and Enßlin, T.A.; Entropy 2021, 23, 1640. doi.org/10.3390/e23121640