Information criteria for non-normalized models

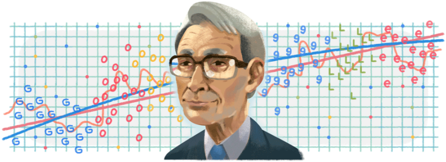
Takeru Matsuda, Masatoshi Uehara, Aapo Hyvärinen

Journal of Machine Learning Research, 2021.

Abstract

- Akaike information criterion (AIC) enables data-driven selection from normalized models.

$$\mathrm{AIC} = -2 \log p(x \mid \hat{\theta}) + 2k$$



  - image from Google (Nov 5, 2017; Akaike's 90th birthday)

- We develop information criteria for non-normalized models.

MLE	Kullback–Leibler divergence	AIC, TIC
score matching	Fisher divergence	SMIC
NCE	Bregman divergence	NCIC

Non-normalized models

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta)$$

$$Z(\theta) = \int \widetilde{p}(x \mid \theta) \mathrm{d}x$$

- Some statistical models are defined by $\widetilde{p}(x \mid \theta)$ and the normalization constant $Z(\theta)$ is computationally intractable
  - e.g. Markov random field, distribution on manifolds

- also known as "energy-based model" in machine learning

$$\widetilde{p}(x \mid \theta) = \exp(-E(x \mid \theta))$$

Estimation methods for non-normalized models

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta)$$

$$Z(\theta) = \int \widetilde{p}(x \mid \theta) \mathrm{d}x$$

- estimate θ from $x_1, \ldots, x_N \sim p(x \mid \theta)$
- MLE is computationally intensive for non-normalized models..
- Several methods have been developed that do not require computation of $Z(\theta)$.
  - pseudo-likelihood (Besag, 1974)
  - contrastive divergence (Hinton, 2002)
  - score matching (Hyvärinen, 2005)
  - noise contrastive estimation (Gutmann and Hyvärinen, 2012)

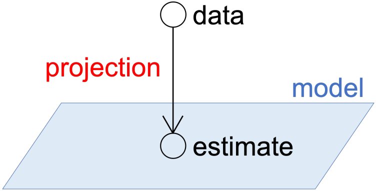
Divergence viewpoint

- divergence (“distance" between probability distributions)

$$D(q,p) \geq 0, \quad D(q,p) = 0 \Leftrightarrow q = p$$

- empirical distribution

$$\hat{q}(x) = \frac{1}{N} \sum_{t=1}^{n} \delta(x - x_t)$$



- projection estimator

$$\hat{\theta}_D = \operatorname*{argmin}_{\theta} D(\hat{q}, p_\theta)$$

MLE	Kullback–Leibler divergence
score matching	Fisher divergence
NCE	Bregman divergence

MLE = KL projection

- Maximum likelihood estimator (MLE)

$$\hat{\theta}_{\mathrm{MLE}} = \operatorname*{argmax}_{\theta} \sum_{t=1}^{N} \log p(x_t \mid \theta)$$

- Kullback-Leibler divergence

$$D_{\mathrm{KL}}(q, p_\theta) = \int q(z) \log \frac{q(z)}{p(z \mid \theta)} \mathrm{d}z$$

- MLE = KL projection

$$\hat{\theta}_{\mathrm{MLE}} = \operatorname*{argmin}_{\theta} \; D_{\mathrm{KL}}(\hat{q}, p_\theta)$$

Score matching
(Hyvärinen, 2005)

Score function

- $q(x)$: probability density on $\mathbb{R}^d$
- score function

$$\nabla_x \log q(x) = \left(\frac{\partial}{\partial x_1} \log q(x), \ldots, \frac{\partial}{\partial x_d} \log q(x) \right)$$

- Important: score function does not involve $Z(\theta)$!!

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta)$$

$$\nabla_x \log p(x \mid \theta) = \nabla_x \log \widetilde{p}(x \mid \theta)$$

Fisher divergence

- Fisher divergence = L^2 distance between score functions

$$D_{\mathrm{F}}(q,p) = \int \|\nabla_x \log q(x) - \nabla_x \log p(x)\|^2 \, q(x)\mathrm{d}x$$

- By using integration by parts in $\mathbb{R}^d$,

$$D_{\mathrm{F}}(q,p) = g(q) + d_{\mathrm{SM}}(q,p)$$

$$d_{\mathrm{SM}}(q,p) = \int \left(2\Delta_x \log p(x) + \|\nabla_x \log p(x)\|^2\right) q(x)\mathrm{d}x$$

Score matching

- Fisher discrepancy from empirical distribution

$$d_{\mathrm{SM}}(\hat{q}, p_\theta) = \frac{1}{N} \sum_{t=1}^{N} \left(2\Delta_x \log \widetilde{p}(x_t \mid \theta) + \|\nabla_x \log \widetilde{p}(x_t \mid \theta)\|^2 \right)$$

- Important: $d_{\mathrm{SM}}(\hat{q}, p_\theta)$ does not involve $Z(\theta)$!!
- score matching estimator

$$\hat{\theta}_{\mathrm{SM}} = \operatorname*{argmin}_{\theta} d_{\mathrm{SM}}(\hat{q}, p_\theta)$$

- This estimator has consistency and asymptotic normality under mild regularity conditions (Hyvärinen, 2005).

Noise contrastive estimation (Gutmann and Hyvärinen, 2012)

Noise contrastive estimation (NCE)

$$\log p(x \mid \theta, c) := \log \widetilde{p}(x \mid \theta) + c, \quad c = -\log Z(\theta)$$

- NCE estimates θ and c simultaneously.
- NCE is based on discrimination between data and noise.
  - similar in spirit to Generative Adversarial Network (GAN)
- In addition to data $x_1, \cdots, x_N \sim p(x \mid \theta)$, we generate noise samples $y_1, \cdots, y_M$ from a noise distribution $n(y)$.
  - should be difficult to discriminate from data

Noise contrastive estimation (NCE)

- The estimate is defined to discriminate between data and noise as accurately as possible.

$$(\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}) = \arg\min_{\theta,c} \ \hat{d}_{\mathrm{NCE}}(\theta, c)$$

$$\begin{aligned}\hat{d}_{\mathrm{NCE}}(\theta, c) = &- \sum_{t=1}^{N} \log \frac{Np(x_t \mid \theta, c)}{Np(x_t \mid \theta, c) + Mn(x_t)} \\ &- \sum_{t=1}^{M} \log \frac{Mn(y_t)}{Np(y_t \mid \theta, c) + Mn(y_t)}\end{aligned}$$

- $\hat{d}_{\mathrm{NCE}}$: negative log-likelihood of the logistic regression classifier
- This estimator has consistency and asymptotic normality under mild regularity conditions (Gutmann and Hyvärinen, 2012).

Bregman divergence induced by NCE

- Gutmann and Hirayama (2011): NCE can be interpreted as projection with respect to a Bregman divergence

$$D_{\mathrm{NCE}}(q, p) = \int d_f\left(\frac{q(x)}{n(x)}, \frac{p(x)}{n(x)}\right) n(x) \mathrm{d}x$$

$$d_f(a, b) = f(a) - f(b) - f'(b)(a - b)$$

$$f(x) = x \log x - \left(\frac{M}{N} + x\right) \log\left(1 + \frac{N}{M} x\right)$$

- Pihlaja et al. (2010) compared other choices of f in simulation
- Uehara et al. (2018): this f minimizes the asymptotic variance

Akaike Information Criterion (AIC) and Takeuchi Information Criterion (TIC)

Setting

$$X_1, \cdots, X_N \sim q(x), \quad N \to \infty$$

- candidate model: $p(x \mid \theta)$
- Maximum Likelihood Estimator (MLE)

$$\hat{\theta}_{\mathrm{MLE}}(x^N) = \arg\max_{\theta} \sum_{t=1}^{N} \log p(x_t \mid \theta)$$

- We want to select a model with smaller KL divergence from the true distribution

$$D_{\mathrm{KL}}(q(z), p(z \mid \hat{\theta}_{\mathrm{MLE}}(x^N)))$$

KL discrepancy and bias correction

- Kullback–Leibler discrepancy

$$d_{\mathrm{KL}}(q, \hat{\theta}_{\mathrm{MLE}}(x^N)) = -\mathrm{E}_q[\log p(z \mid \hat{\theta}_{\mathrm{MLE}}(x^N))]$$

is equivalent to Kullback–Leibler divergence (up to constant)

$$D_{\mathrm{KL}}(q, \hat{\theta}_{\mathrm{MLE}}(x^N)) = \mathrm{E}_q[\log q(z)] + d_{\mathrm{KL}}(q, \hat{\theta}_{\mathrm{MLE}}(x^N))$$

$\rightarrow$ We estimate expected KL discrepancy for model selection

- The estimate

$$d_{\mathrm{KL}}(\hat{q}, \hat{\theta}_{\mathrm{MLE}}(x^N)) = -\frac{1}{N} \sum_{t=1}^{N} \log p(x_t \mid \hat{\theta}_{\mathrm{MLE}}(x^N))$$

has negative bias becuase it uses data twice.

$\rightarrow$ We correct its bias

Akaike Information Criterion

- Akaike Information Criterion (AIC; Akaike, 1974)

$$\mathrm{AIC} = -2\sum_{t=1}^{N} \log p(x_t \mid \hat{\theta}_{\mathrm{MLE}}(x)) + 2 \cdot \dim(\theta)$$

  - second term: bias correction

Proposition

If the model is well-specified ($q(x) = p(x \mid \theta^*)$), then AIC is an approximately unbiased estimator of the expected KL discrepancy:

$$\mathrm{E}_\theta[\mathrm{AIC}] = -2N\mathrm{E}_\theta[\log p(z \mid \hat{\theta}_{\mathrm{MLE}}(x))] + O(N^{-1})$$

Takeuchi Information Criterion (TIC)

- How about mis-specified case?
- Takeuchi Information Criterion (TIC)

$$\mathrm{TIC} = -2\sum_{t=1}^{N} \log p(x_t \mid \hat{\theta}_{\mathrm{MLE}}(x)) + 2\mathrm{tr}(\hat{I}\hat{J}^{-1})$$

$$\hat{I}_{ij} = \frac{1}{N}\sum_{t=1}^{N} \frac{\partial}{\partial\theta_i}\log p(x_t \mid \theta)\frac{\partial}{\partial\theta_j}\log p(x_t \mid \theta)\bigg|_{\theta=\hat{\theta}_{\mathrm{MLE}}(x^N)}$$

$$\hat{J}_{ij} = -\frac{1}{N}\sum_{t=1}^{N} \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x_t \mid \theta)\bigg|_{\theta=\hat{\theta}_{\mathrm{MLE}}(x^N)}$$

Proposition

$$\mathrm{E}_q[\mathrm{TIC}] = -2N\mathrm{E}_q[\log p(z \mid \hat{\theta}_{\mathrm{MLE}}(x))] + o(1)$$

Information criterion for NCE

Recall: NCE and Bregman divergence

- Bregman divergence

$$D_{\mathrm{NCE}}(q, p) = g(q) + d_{\mathrm{NCE}}(q, p)$$

- NCE

$$(\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}) = \operatorname*{argmin}_{\theta, c} d_{\mathrm{NCE}}(\hat{q}, p_{\theta,c})$$

$$\begin{aligned} d_{\mathrm{NCE}}(\hat{q}, p_{\theta,c}) = &- \sum_{t=1}^{N} \log \frac{Np(x_t \mid \theta, c)}{Np(x_t \mid \theta, c) + Mn(x_t)} \\ &- \sum_{t=1}^{M} \log \frac{Mn(y_t)}{Np(y_t \mid \theta, c) + Mn(y_t)} \end{aligned}$$

Information criterion for NCE (general case)

$$X_1, \dots, X_N \sim q(x), \quad Y_1, \dots, Y_M \sim n(y), \quad M/N \to \nu$$

Theorem 1

The quantity

$$\mathrm{NCIC}_1 = N d_{\mathrm{NCE}}(\hat{q}, \hat{p}) + \mathrm{tr}(\hat{I}\hat{J}^{-1})$$

is an approximately unbiased estimator of $N \mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q, \hat{p})]$:

$$\mathrm{E}_{x,y}[\mathrm{NCIC}_1] = N \mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q, \hat{p})] + o(1)$$

- proof: asymptotics for stratified sampling (Wooldridge, 2001)
  - two strata: data (size N) and noise (size M)

Information criterion for NCE (well-specified case)

$$\hat{b}(z) = \frac{\hat{p}(z)n(z)}{\hat{r}(z)^2}, \quad \hat{r}(z) = \frac{N}{N+M}\hat{p}(z) + \frac{M}{N+M}n(z)$$

Theorem 2

If the model is well-specified ($q(x) = p(x \mid \xi^*)$), then the quantity

$$\mathrm{NCIC}_2 = N d_{\mathrm{NCE}}(\hat{q}, \hat{p}) + m - \frac{1}{N+M}\left(\sum_{t=1}^{N} \hat{b}(x_t) + \sum_{t=1}^{M} \hat{b}(y_t)\right)$$

is an approximately unbiased estimator of $N \mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q, \hat{p})]$:

$$\mathrm{E}_{x,y}[\mathrm{NCIC}_2] = N \mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q, \hat{p})] + o(1)$$

- easier to compute than NCIC_1

Bias correction in NCIC

- non-normalized model ($m = 3$ parameters)

$$p(x \mid \theta, c) = \exp(\theta_1 x^2 + \theta_2 x + c)$$

- data ($N = 10^3$): $(1 - \varepsilon) \cdot \mathrm{N}(0, 1) + \varepsilon \cdot \mathrm{N}(0, 10)$ (Gaussian mixture)
  - When $\varepsilon = 0$, the model is well-specified.

- noise ($M = 10^3$): $\mathrm{N}(0, 1)$

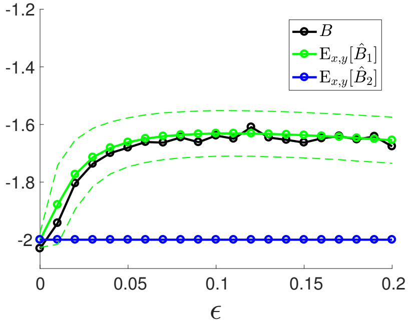
Bias correction in NCIC

- true bias

$$B = N\mathrm{E}_{x,y}\left[\hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}})\right] - N\mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q, \hat{p})]$$

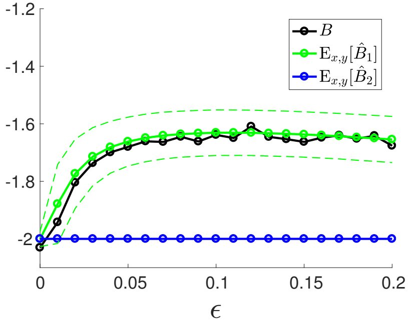
- bias estimate for NCIC_1 and NCIC_2

$$\hat{B}_1 = -\mathrm{tr}(\hat{I}\hat{J}^{-1}), \quad \hat{B}_2 = -m + \frac{1}{N+M}\left(\sum_{t=1}^{N}\hat{b}(x_t) + \sum_{t=1}^{M}\hat{b}(y_t)\right)$$



Bias correction in NCIC

- When $\varepsilon = 0$ (well-specified case), $B \approx -(m-1) = -2$ and both $\mathrm{E}_{x,y}[\hat{B}_1]$ and $\mathrm{E}_{x,y}[\hat{B}_2]$ are close to this value.
- When $\varepsilon > 0$ (mis-specified case), B and $\mathrm{E}_{x,y}[\hat{B}_1]$ coincide well.
- NCIC_2 has much smaller variance than NCIC_1.
  - Also, NCIC_2 is easier to compute than NCIC_1.
  - similar to TIC and AIC



Information criterion for score matching

Recall: Score matching and Fisher divergence

- Fisher divergence

$$D_{\mathrm{F}}(q,p) = g(q) + d_{\mathrm{SM}}(q,p)$$

$$d_{\mathrm{SM}}(q,p) = \int \left(2\Delta_x \log p(x) + \|\nabla_x \log p(x)\|^2\right) q(x)\mathrm{d}x$$

- score matching estimator

$$\hat{\theta}_{\mathrm{SM}} = \underset{\theta}{\operatorname{argmin}} \; d_{\mathrm{SM}}(\hat{q}, p_\theta) = \frac{1}{N}\sum_{t=1}^{N} \rho_{\mathrm{SM}}(x_t, \theta)$$

$$\rho_{\mathrm{SM}}(x,\theta) = 2\Delta_x \log \widetilde{p}(x \mid \theta) + \|\nabla_x \log \widetilde{p}(x \mid \theta)\|^2$$

Information criterion for score matching

$$\hat{I} = \frac{1}{N}\sum_{t=1}^{N} \nabla_\theta \rho_{\mathrm{SM}}(x_t, \theta) \nabla_\theta \rho_{\mathrm{SM}}(x_t, \theta)^\top \Big|_{\theta=\hat{\theta}}$$

$$\hat{J} = \frac{1}{N}\sum_{t=1}^{N} \nabla_\theta^2 \rho_{\mathrm{SM}}(x_t, \theta) \Big|_{\theta=\hat{\theta}}$$

Theorem 3

The quantity

$$\mathrm{SMIC} = N d_{\mathrm{SM}}(\hat{q}, \hat{p}) + \mathrm{tr}(\hat{I}\hat{J}^{-1})$$

is an approximately unbiased estimator of $N\mathrm{E}_q[d_{\mathrm{SM}}(q, \hat{p})]$:

$$\mathrm{E}_x[\mathrm{SMIC}] = N\mathrm{E}_x[d_{\mathrm{SM}}(q, \hat{p})] + o(1)$$

Bias correction in SMIC

- model: $\mathrm{N}(\mu, \sigma^2)$
- data ($N = 10^3$): $(1 - \varepsilon) \cdot \mathrm{N}(0, 1) + \varepsilon \cdot \mathrm{N}(0, 10)$ (Gaussian mixture)
- true bias

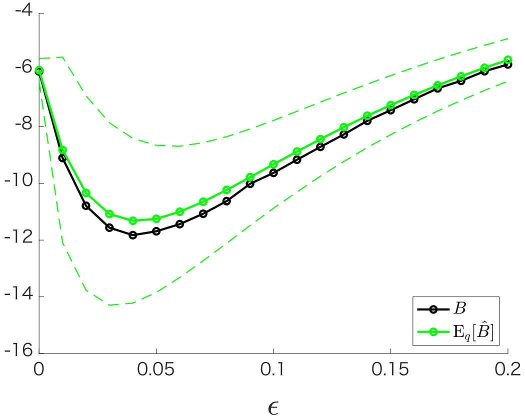
$$B = N\mathrm{E}_q\left[\hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}})\right] - N\mathrm{E}_q[d_{\mathrm{SM}}(q, \hat{p})]$$

- bias estimate in SMIC

$$\hat{B} = -\mathrm{tr}(\hat{I}\hat{J}^{-1})$$

Bias correction in SMIC

- Consistent with Theorem 3, B and $\mathrm{E}_{x,y}[\hat{B}]$ coincide quite well.
  - The bias is larger than NCE.



Applications

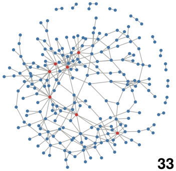
Truncated Gaussian graphical model

truncated Gaussian graphical model (Lin et al., 2016)

$$p(x \mid \Sigma) \propto \exp\left(-\frac{1}{2}x^\top \Sigma^{-1} x\right), \quad x \in \mathbb{R}_+^d$$

- $G = (V, E)$: undirected graph with $V = \{1, \cdots, d\}$
- $\Sigma \succ 0$, $(\Sigma^{-1})_{ij} = 0$ if $(i, j) \notin E$ (no edge between i and j)

- The normalization constant is computationally intractable.
- Lin et al. (2016) estimated this model by l_1-regularized score matching.
  - equivalent to LASSO mathematically
  - application: RNAseq data



Edge selection performance

$$\Sigma^{-1} = \begin{pmatrix} 1 & \sigma^{12} & 0 \\ \sigma^{12} & 1 & 0.55 \\ 0 & 0.55 & 1 \end{pmatrix}$$

- counts of selection of each edge over 100 simulations
  - NCIC_1 / NCIC_2 / SMIC

$N = M = 100$

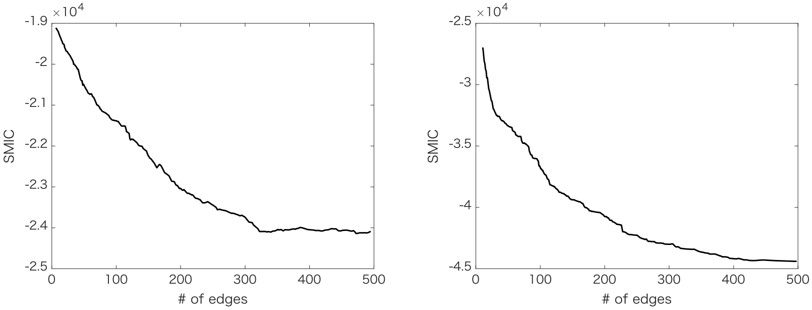
σ^{12}	(1,2)	(1,3)	(2,3)
0.2	26/20/37	22/17/30	45/39/58
0.3	38/27/44	20/19/25	60/59/71
0.5	56/49/62	23/17/33	42/39/62

$N = M = 1000$

σ^{12}	(1,2)	(1,3)	(2,3)
0.2	63/59/59	14/13/18	100/100/100
0.3	88/88/89	20/15/18	100/99/100
0.5	97/96/98	15/14/22	99/99/99

Application to RNAseq data

- RNAseq data for 40 genes
  - used in Lin et al. (2016)
- SMIC w.r.t. edge counts
  - left: truncated GGM, right: log-GGM



- Log-GGM has better fit to RNAseq data in this case.

Overcomplete independent component analysis (ICA)

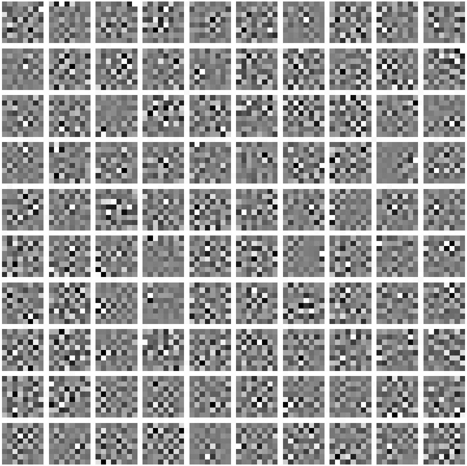
energy-based overcomplete ICA model (Teh et al., 2004)

$$p(x) \propto \exp\left(\sum_{b=1}^{B} G(w_b^\top x)\right), \quad G(u) = -|u|$$

- data ($N = 5 \times 10^4$): 8 × 8 image patches from natural images
  - analyzed in Hyvärinen (2005) with score matching
- noise ($M = 5 \times 10^4$): Gaussian with the same mean and covariance as data

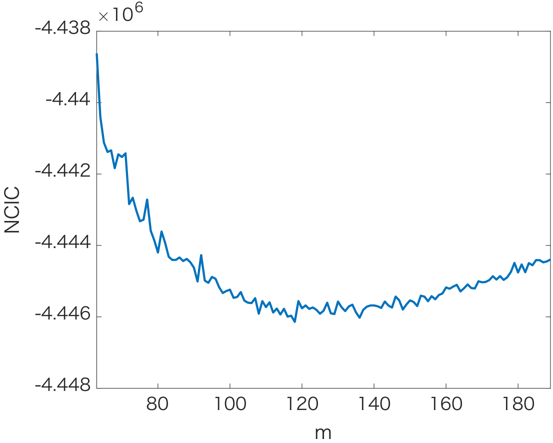
- We select the number of filters B by minimizing NCIC_2.

Data



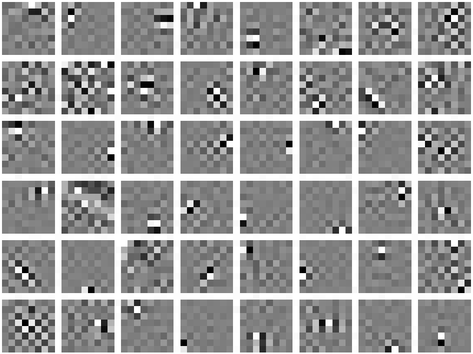
Model selection result

- NCIC_2 takes minimum at $B = 118$.
  - Hyvärinen (2005) set $B = 200$.



Filters

- estimated filters w_b when $B = 118$



- They respond to localized patterns (like V1 neurons).
  - similar to the filters obtained in Hyvärinen (2005).

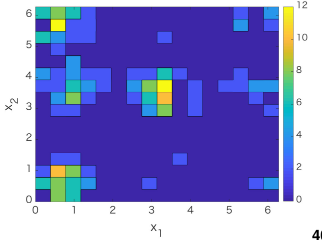
Directional data analysis

Bivariate von Mises distribution (Singh et al., 2002)

$$p(x_1, x_2 \mid \theta) \propto \exp(\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2) + \lambda_{12} \sin(x_1 - \mu_1) \sin(x_2 - \mu_2)), \ (x_1, x_2) \in [0, 2\pi)^2$$

- $\theta = (\kappa_1, \kappa_2, \mu_1, \mu_2, \lambda_{12})$ with $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, $\mu_1, \mu_2 \in [0, 2\pi)$
- The normalization constant is computationally intractable (infinite sum of Bessel functions)

- daily wind direction at Tokyo in 2018, 00:00 (x_1) & 12:00 (x_2)
- NCIC comparison ($-1941 < -1919$) implies that x_1 and x_2 are dependent ($\lambda_{12} \neq 0$)



Summary

- We developed information criteria for non-normalized models estimated by NCE or score matching.

MLE	KL divergence	AIC, TIC
score matching	Fisher divergence	SMIC
NCE	Bregman divergence	NCIC

- By using these criteria, we can select the appropriate non-normalized model in a data-driven manner.

- paper: M., Uehara and Hyvärinen. *Journal of Machine Learning Research*, 2021.