Machine Learning and
Many-Body Physics
@KITS, UCAS, Beijing
7 July, 2017

Bayesian spectral deconvolution: How many peaks are there in this spectrum?

Satoru Tokuda
MathAM-OIL, AIST

ST, K. Nagata and M. Okada, *J. Phys. Soc. Jpn*. 86, 024001, 2017.
(arXiv:1607.07590)

Collaborators

Kenji Nagata
AIRC, AIST

Masato Okada
UTokyo/NIMS

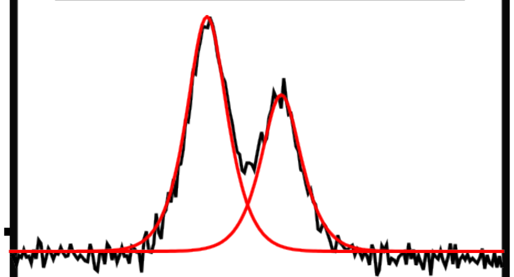
ST, K. Nagata and M. Okada, *J. Phys. Soc. Jpn*. 86, 024001, 2017. (arXiv:1607.07590)

Spectroscopy

Materials science

Life science

Spectroscopy



Planetary science

Some pictures have been removed from the original version due to copyright.

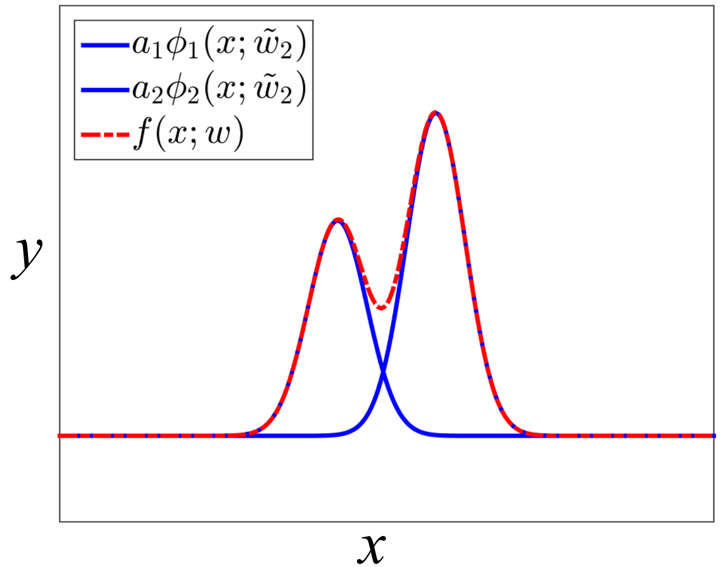
Effective model of spectrum

$$f(x;w)=\sum_{k=1}^{K} a_k \phi_k(x;\tilde{w}_k), \quad w=\{a_k,\tilde{w}_k\}_{k=1}^{K}$$

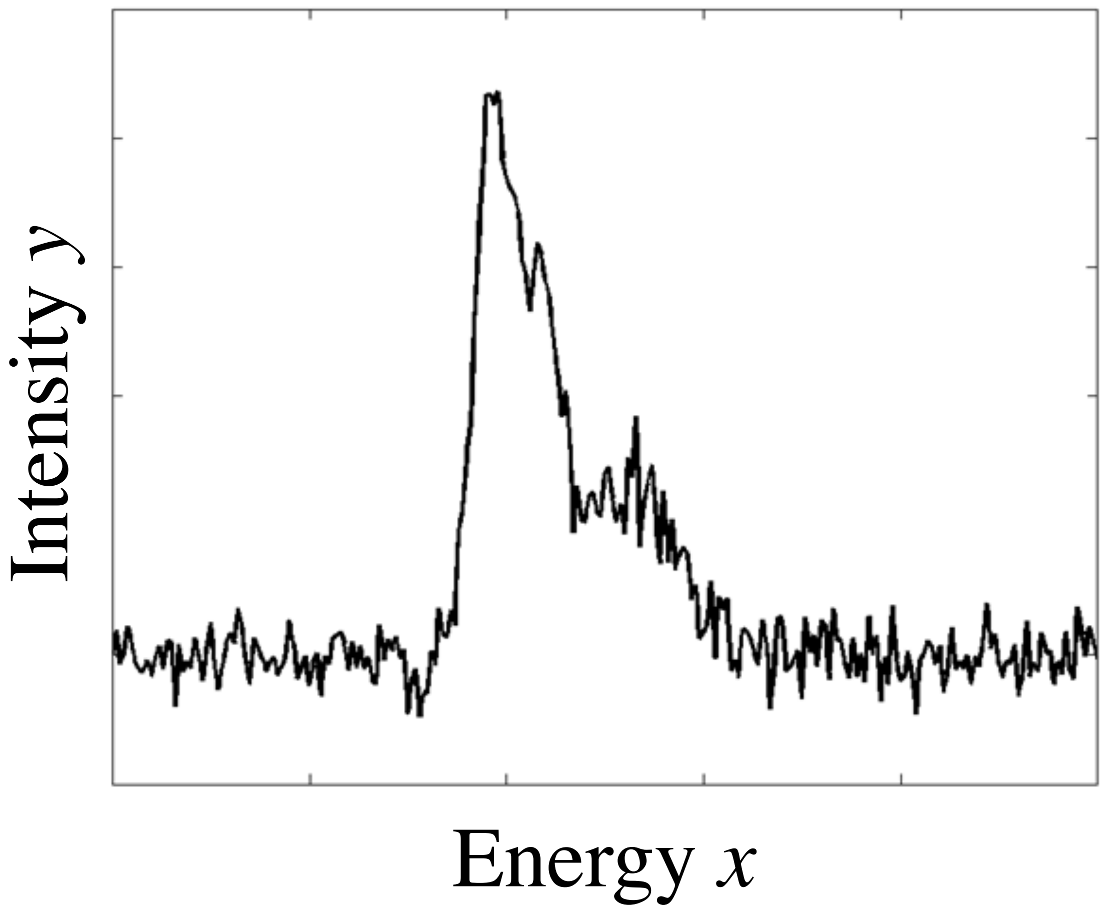
$$\phi_k(x;\tilde{w}_k)=\begin{cases}\exp\left(-\dfrac{(x-\mu_k)^2}{2\rho_k^2}\right) & \text{Thermal Doppler broadening}\\ \dfrac{1}{(x-\mu_k)^2+\gamma_k^2} & \text{Natural broadening}\end{cases}$$

Physical parameters

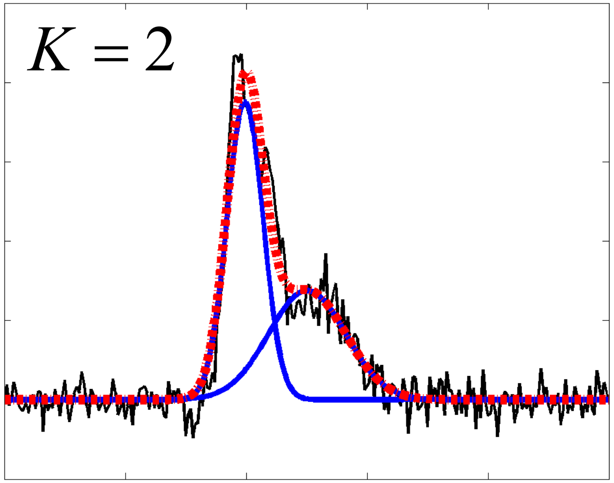
K: Peak number
μ_k: Energy level
a_k: Number density
ϱ_k: Temperature
γ_k: Lifetime

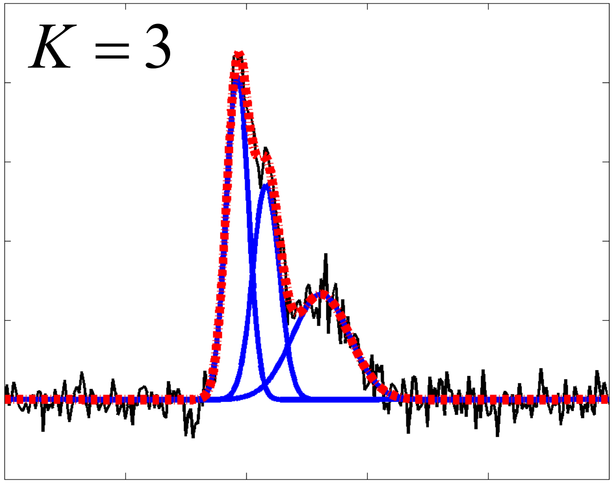


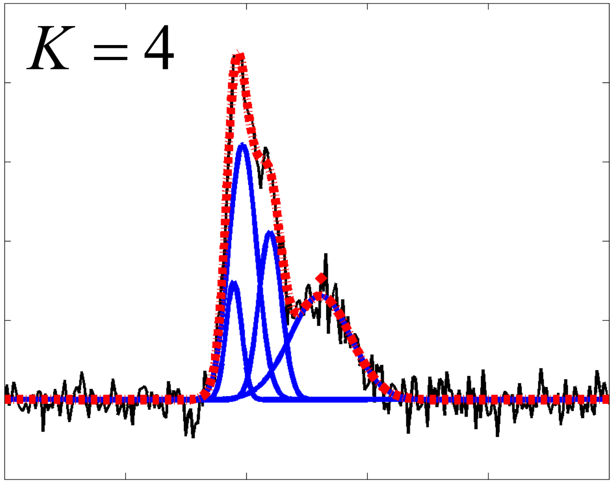
How many peaks are there?



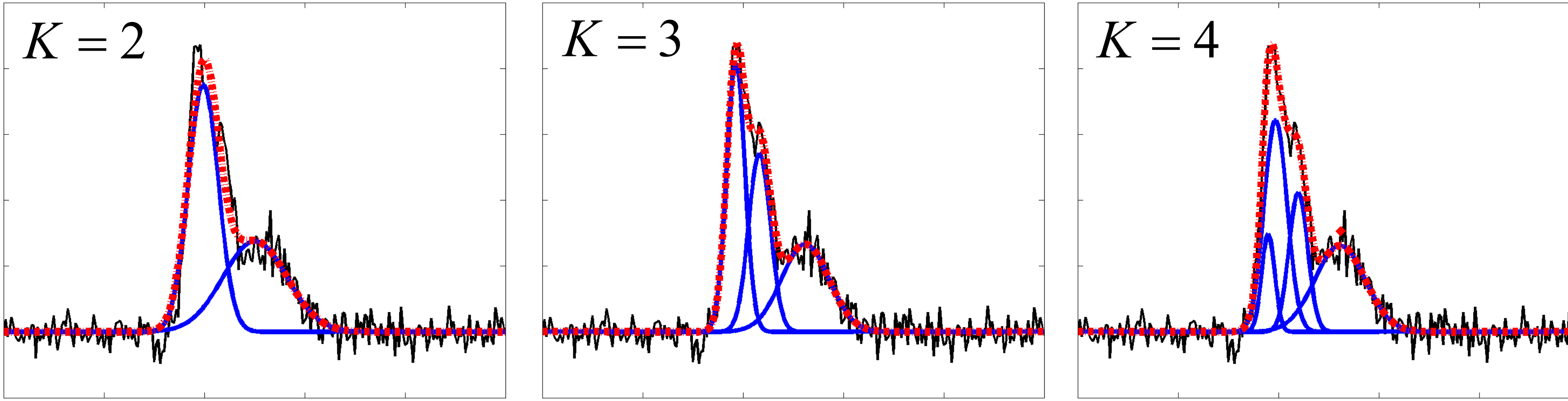
Peak fitting: Learning in radial basis function network





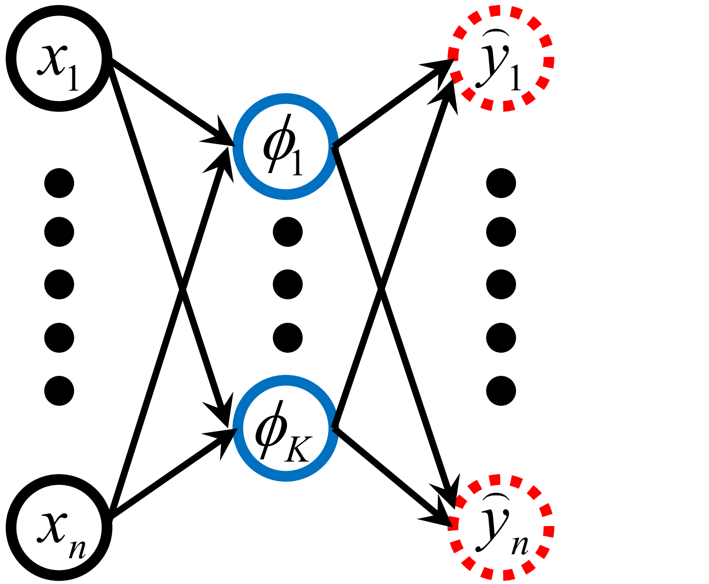


Peak fitting: Learning in radial basis function network



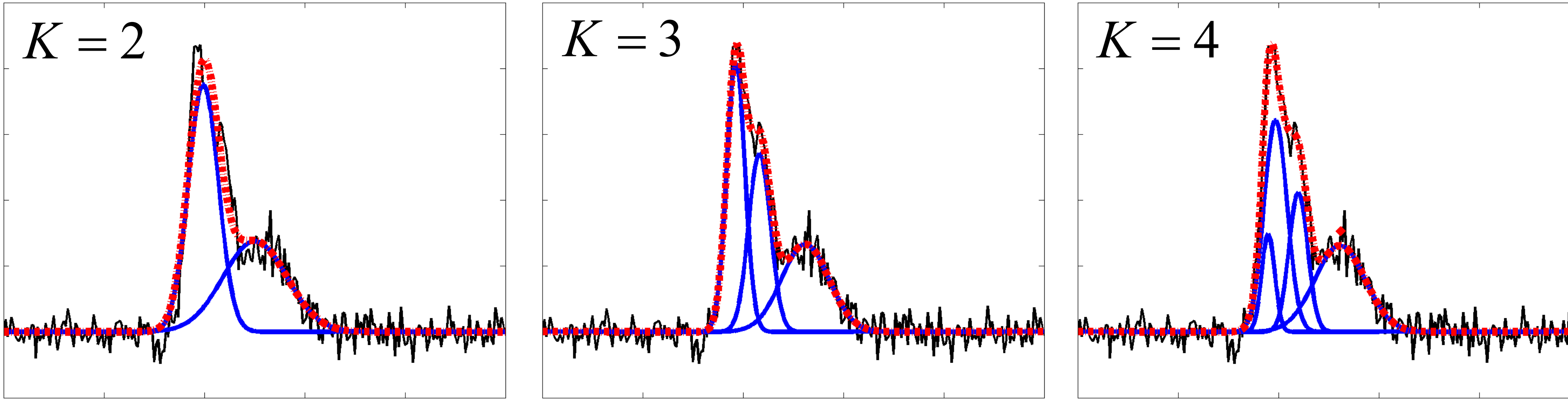
RBF network : $\boxed{f(x;w)} = \sum_{k=1}^{K} a_k \boxed{\exp\left(-\frac{(x-\mu_k)^2}{2\rho_k^2}\right)}$

$\widehat{y}_i$ $\qquad$ $\phi_k(x_i;\mu_k,\rho_k^2)$



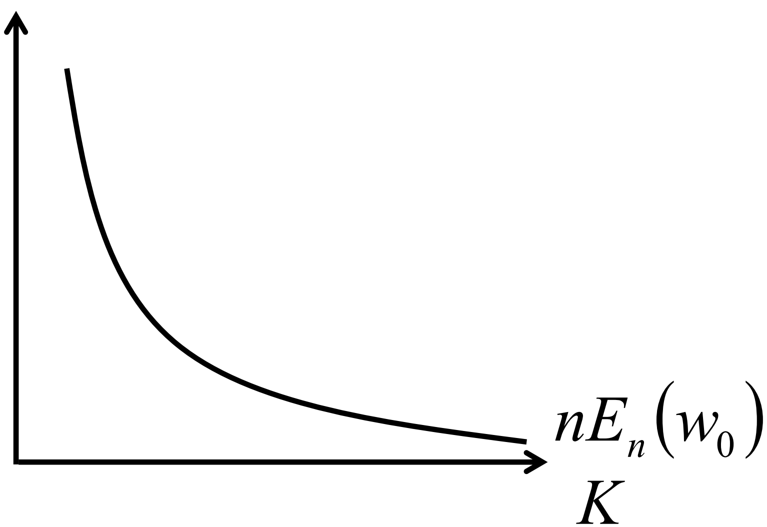
Not to predict the spectral curve $\widehat{y}_i$
But to identify the peaks ϕ_k themselves

Peak fitting: Learning in radial basis function network



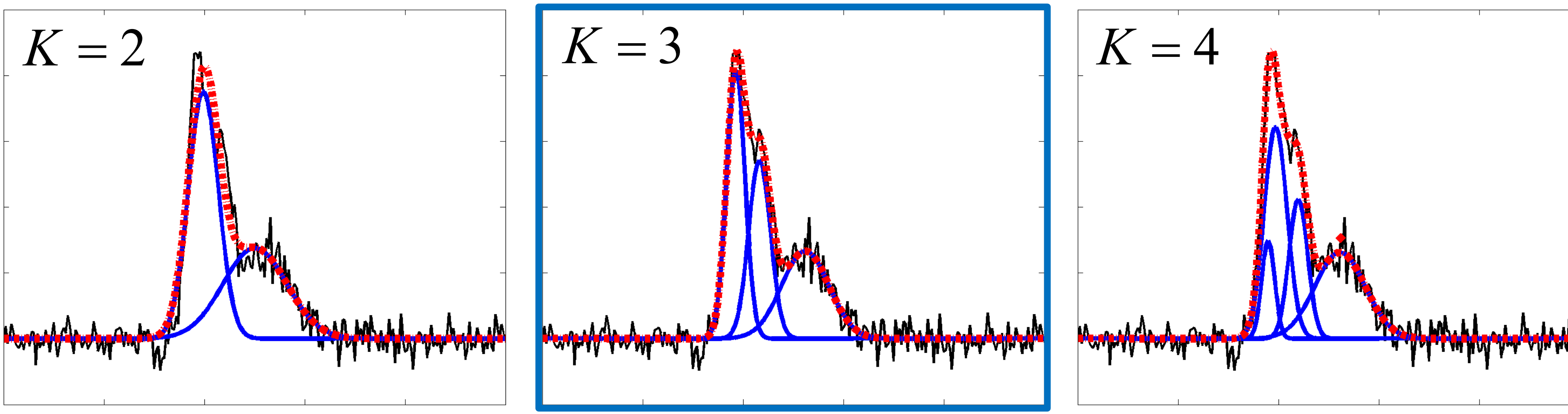
RBF network : $f(x;w)=\sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\rho_k^2}\right)$

Square error: $E_n(w)=\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-f(X_i;w)\right)^2$



Bayesian spectral deconvolution

[Nagata et al., *Neural Netw.*, 2012]

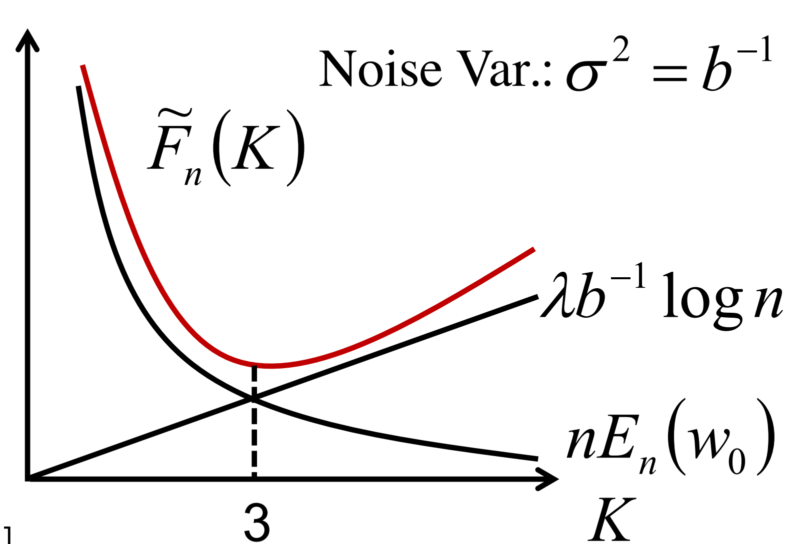


RBF network : $f(x;w)=\sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\rho_k^2}\right)$

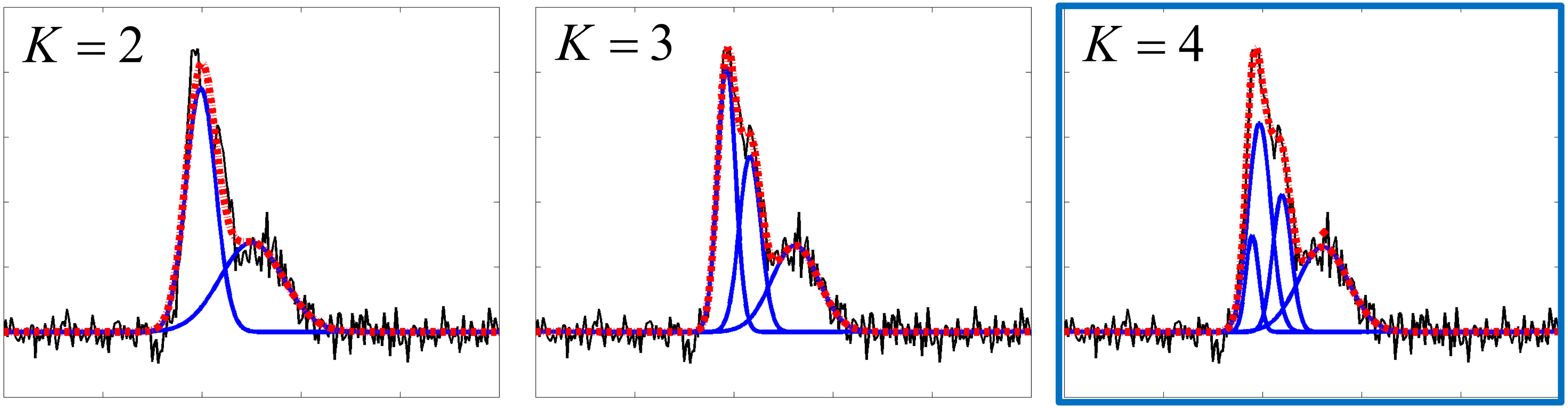
Square error: $E_n(w)=\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-f(X_i;w)\right)^2$

Evaluation function : $\widetilde{F}_n(K)=nE(w_0)+\lambda b^{-1}\log n$

[Watanabe, *Neural Compt.*, 2001]



Overfitting: Underrate unknown noise

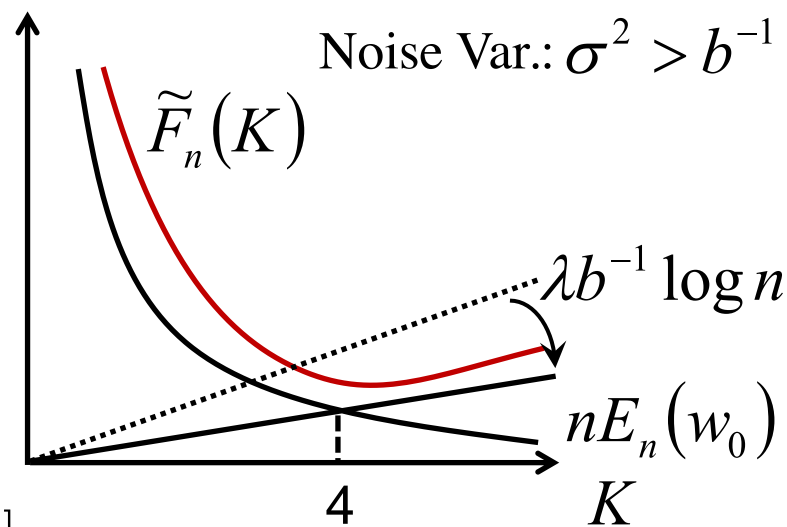


RBF network : $f(x;w)=\sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\rho_k^2}\right)$

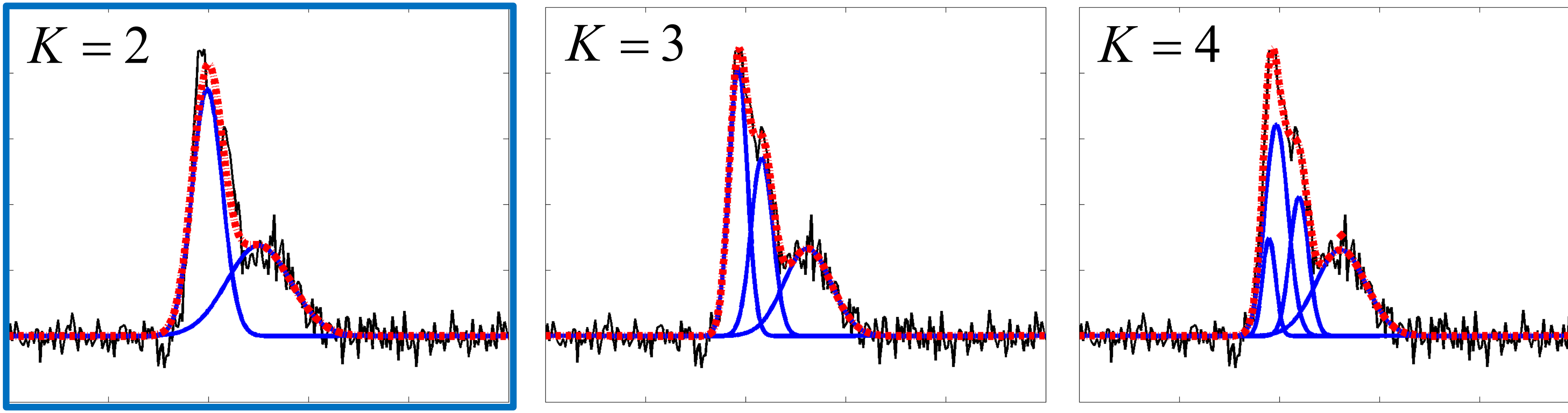
Square error: $E_n(w)=\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - f(X_i;w)\right)^2$

Evaluation function : $\widetilde{F}_n(K)=nE(w_0)+\lambda b^{-1}\log n$

[Watanabe, *Neural Compt.*, 2001]



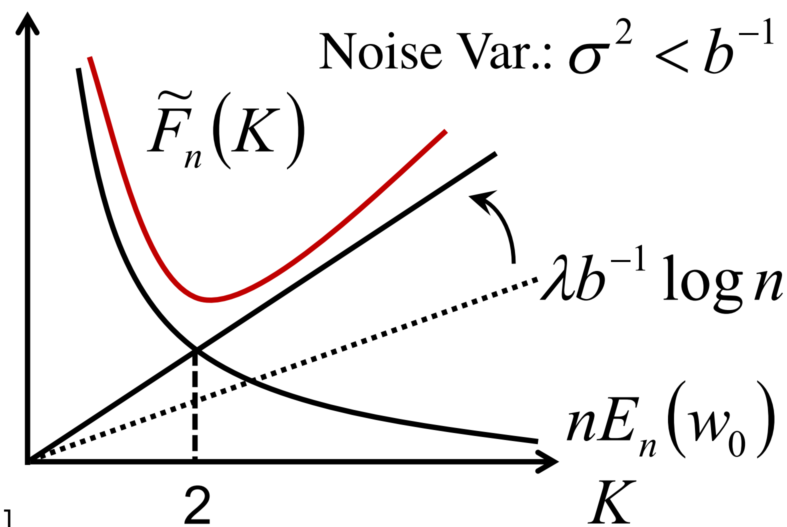
Underfitting: Overrate unknown noise



RBF network : $f(x;w)=\sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\rho_k^2}\right)$

Square error: $E_n(w)=\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - f(X_i;w)\right)^2$

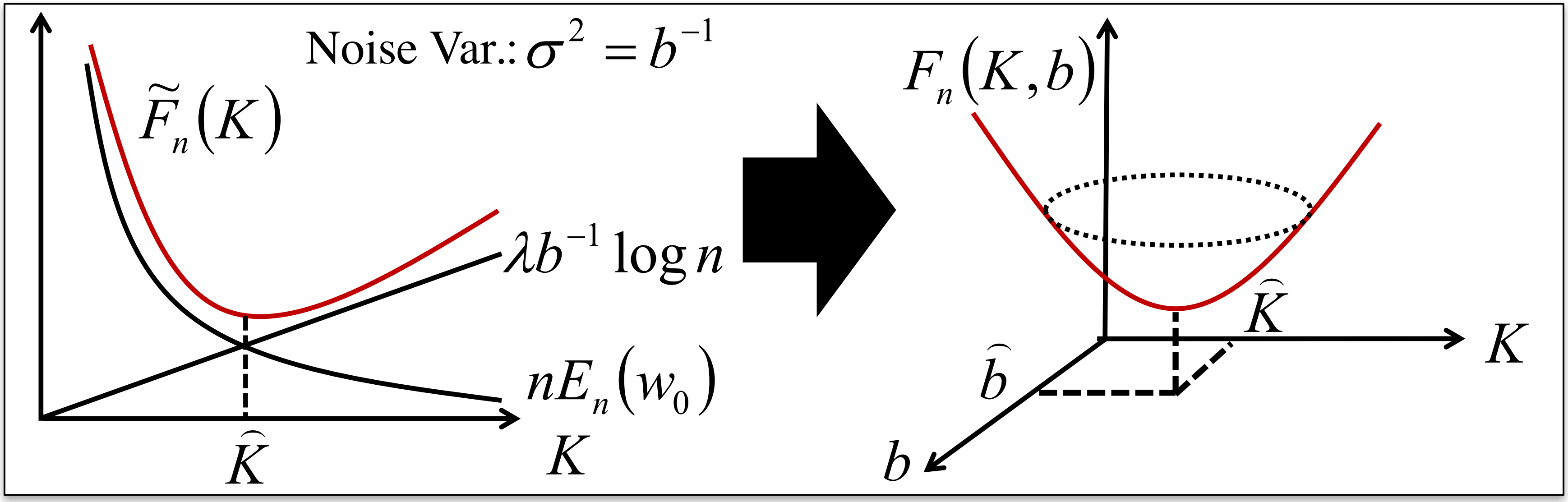
Evaluation function : $\tilde{F}_n(K)=nE(w_0)+\lambda b^{-1}\log n$



[Watanabe, *Neural Compt.*, 2001]

Aim of this study

- Propose a framework that enables the joint estimate of peak number and noise variance from the observed spectrum by modifying the previous framework of Bayesian spectral deconvolution



ST, K. Nagata and M. Okada, *J. Phys. Soc. Jpn*. 86, 024001, 2017. (arXiv:1607.07590)

Model:

The forward problem

Effective model of spectrum

$$f(x;w)=\sum_{k=1}^{K} a_k \phi_k(x;\tilde{w}_k), \quad w=\{a_k,\tilde{w}_k\}_{k=1}^{K}$$

$$\phi_k(x;\tilde{w}_k)=\begin{cases} \exp\left(-\dfrac{(x-\mu_k)^2}{2\rho_k^2}\right) & \text{Thermal Doppler broadening} \\ \dfrac{1}{(x-\mu_k)^2+\gamma_k^{\,2}} & \text{Natural broadening} \end{cases}$$

Physical parameters

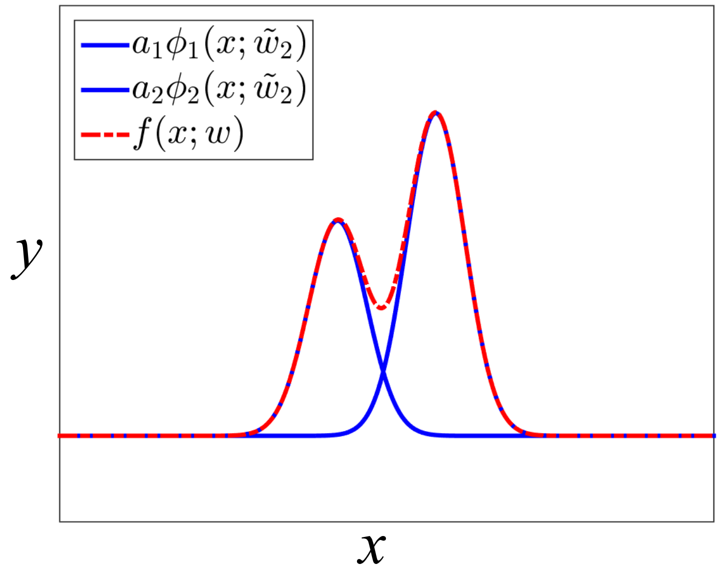
K: Peak number

μ_k: Energy level

a_k: Number density

ρ_k: Temperature

γ_k: Lifetime



Statistical model of spectroscopy

$D = \{X_i, Y_i\}_{i=1}^{n}$ Measured data

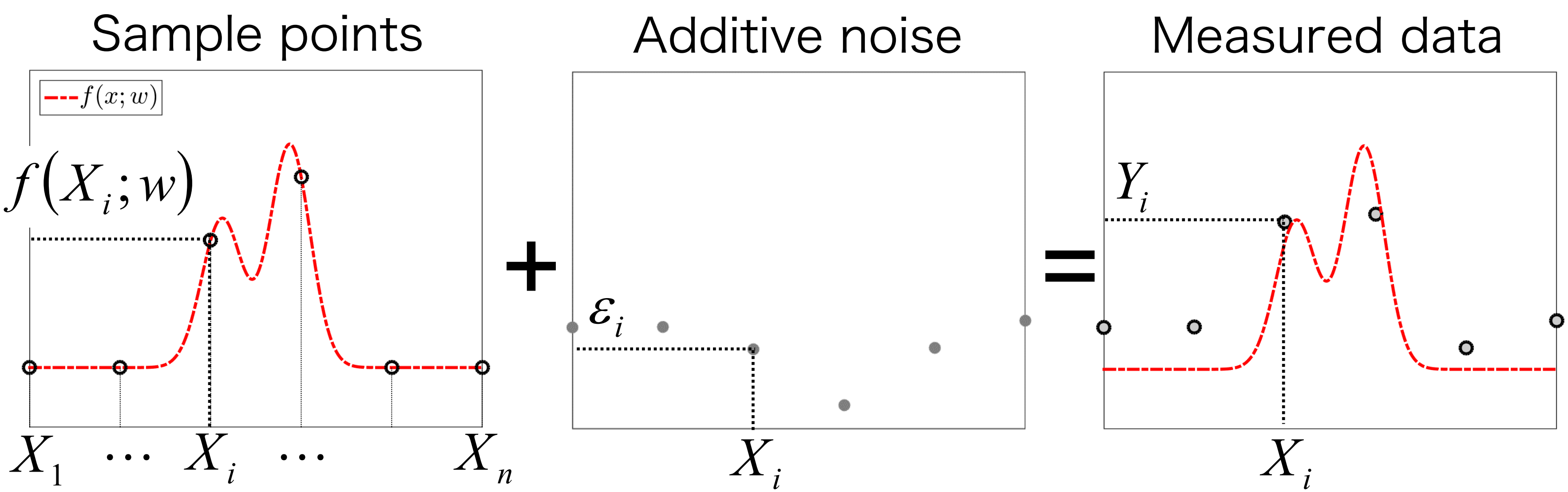
$X_i = X_{i-1} + \Delta x \quad (\Delta x > 0)$ Sample points

$Y_i = f(X_i; w) + \varepsilon_i, \quad \varepsilon_i \sim N\left(f(X_i; w), b^{-1}\right)$ Additive white Gaussian noise

$\boxed{\text{Measurement parameters}}$

b^{-1}: Noise variance

Δx: Energy resolution



Statistical model of spectroscopy

$D = \{X_i, Y_i\}_{i=1}^{n}$ Measured data

$X_i = X_{i-1} + \Delta x \quad (\Delta x > 0)$ Sample points

$Y_i = f(X_i; w) + \varepsilon_i, \quad \varepsilon_i \sim N\left(f(X_i; w), b^{-1}\right)$ Additive white Gaussian noise

$\boxed{\text{Measurement parameters}}$

b^{-1}: Noise variance

Δx: Energy resolution

Statistical model: $p(Y_i | X_i; w) := \sqrt{\frac{b}{2\pi}} \exp\left(-\frac{b}{2} (Y_i - f(X_i; w))^2 \right)$

Likelihood: $p(Y^n | X^n, w, b) := \prod_{i=1}^{n} p(Y_i | X_i, w, b) = \left(\frac{b}{2\pi} \right)^{\frac{n}{2}} \exp(-nbE_n(w))$

Square error: $E_n(w) := \frac{1}{2n} \sum_{i=1}^{n} (Y_i - f(X_i; w))^2$

Bayesian formulation:

The inverse problem

Bayesian inference

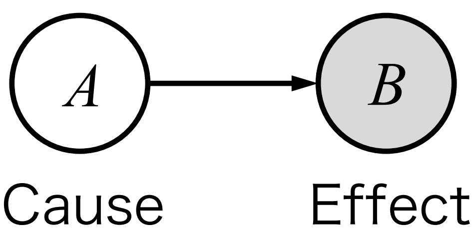
Bayes' theorem

$$\boxed{p(A|B)=\frac{p(B|A)p(A)}{p(B)}}$$

$$\because\ p(A,B)=p(A|B)p(B)=p(B|A)p(A)$$

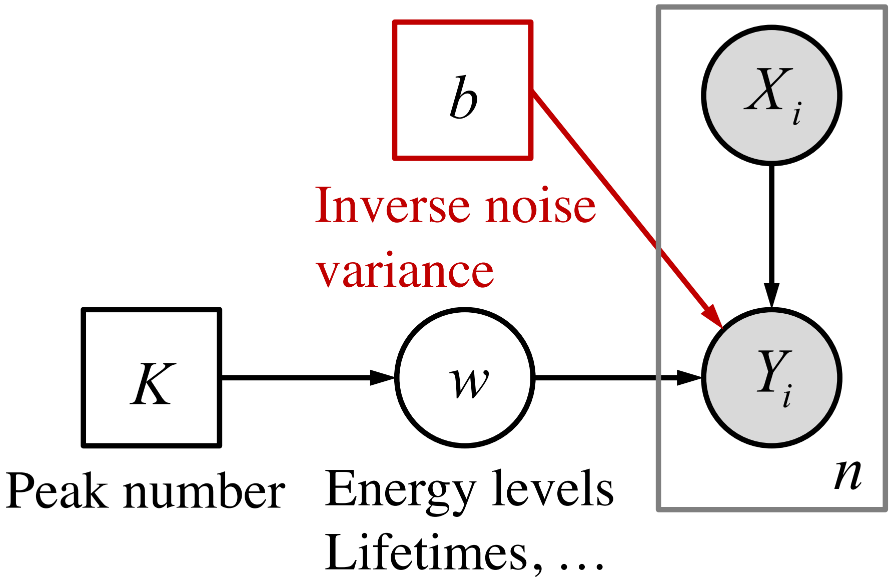
$$p(B)=\begin{cases}\sum_A p(B|A)p(A) & \text{(as } A \text{ is discrete-valued)}\\ \int dA\, p(B|A)p(A) & \text{(as } A \text{ is continuous-valued)}\end{cases}$$

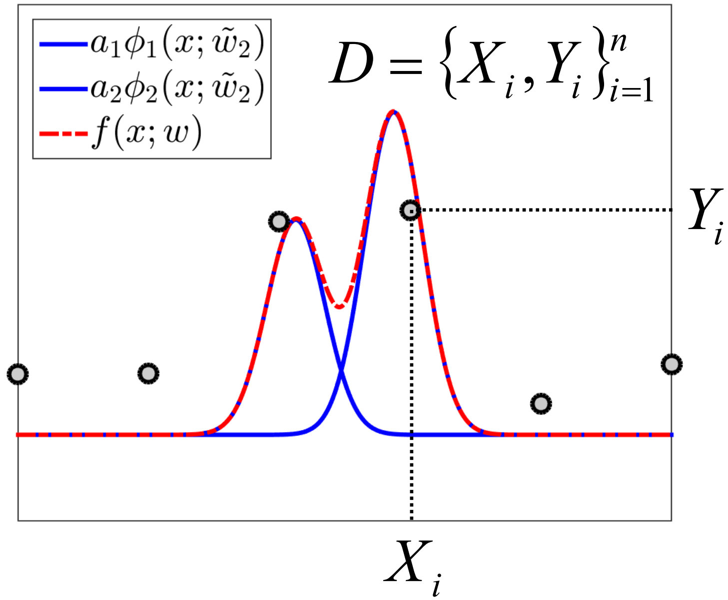
Not only the effect B but also the cause A is treated as random variable to estimate $p(A|B)$



Trace back to the cause issuing from its effect

Bayesian hierarchical modeling





[1] Regression

$$p(w|D,K,b)=\frac{p(Y^n|X^n,w,b)p(w|K)}{p(Y^n|X^n,K,b)}$$

$$p(Y^n|X^n,w,b)=\left(\frac{b}{2\pi}\right)^{n/2}\exp(-nbE_n(w))$$

$$p(Y^n|X^n,K,b)=\int dw\,p(Y^n|X^n,w,b)p(w|K)$$

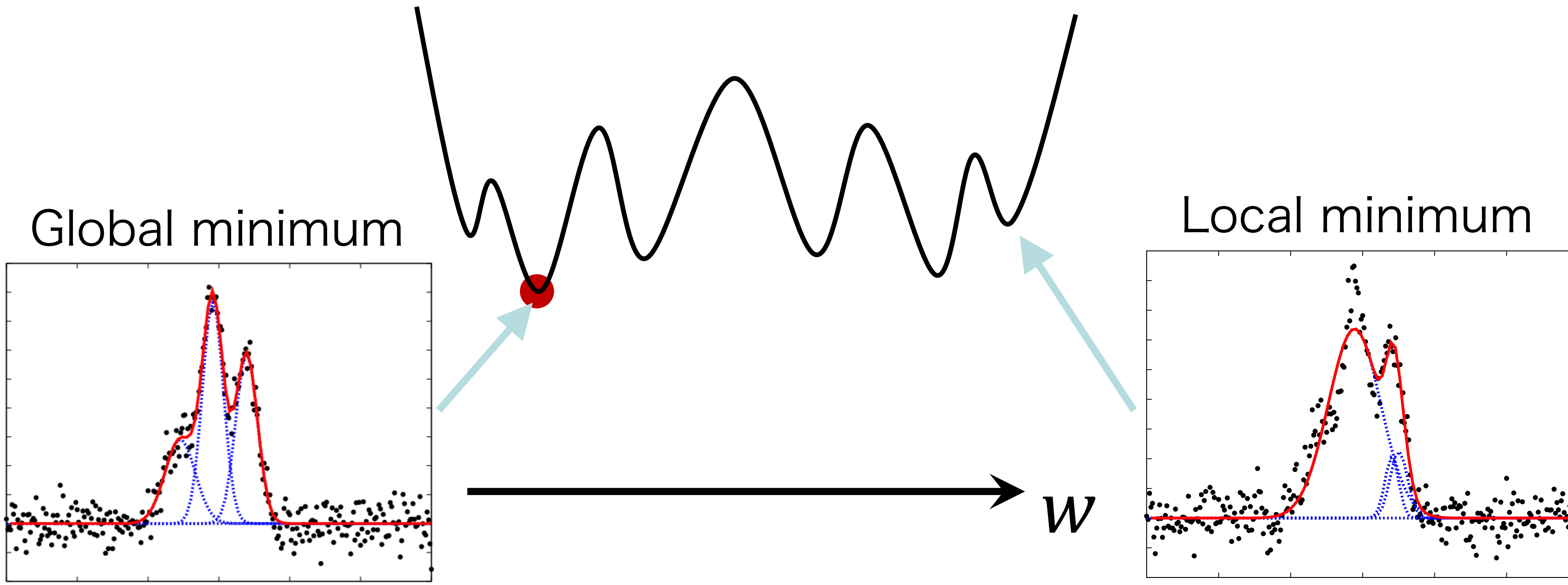
[2] Model Selection

$$p(K,b|D)=\frac{p(Y^n|X^n,K,b)p(K)p(b)}{p(Y^n|X^n)}$$

$$p(Y^n|X^n)=\sum_K\int db\,p(Y^n|X^n,K,b)p(K)p(b)$$

$$(\hat{K},\hat{b}):=\arg\max_{K,b}p(Y^n|X^n,K,b)$$

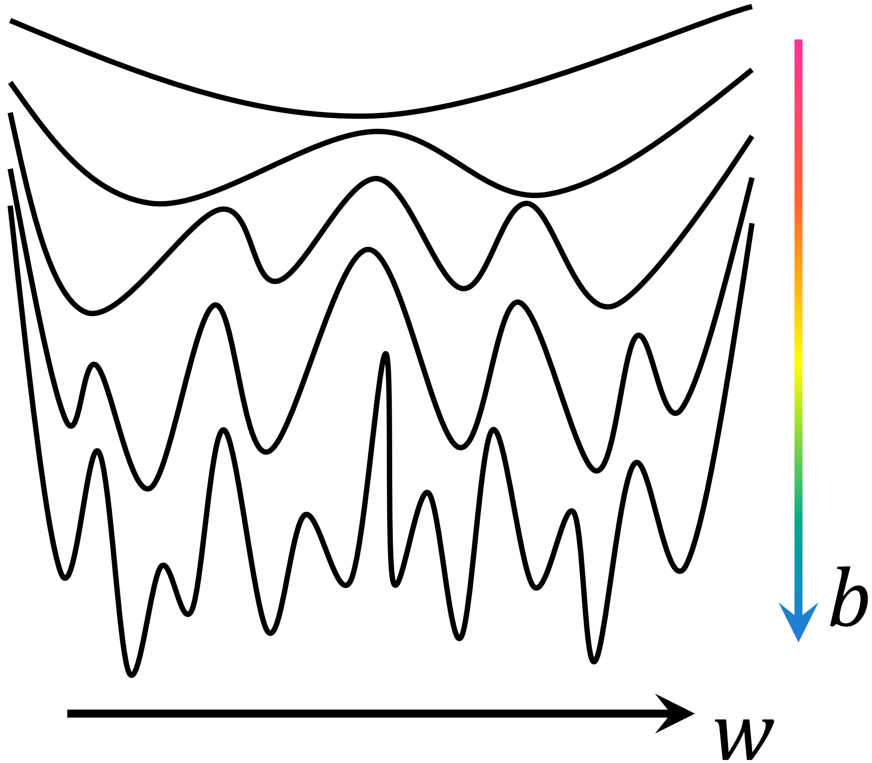
Posterior distribution



$$p(w|D,K,b)=\frac{p(Y^n|X^n,w,b)p(w|K)}{p(Y^n|X^n,K,b)}=\frac{1}{\tilde{Z}_n(K,b)}\exp(-nbE_n(w))\varphi(w|K)$$

Solution space including least-squares solution

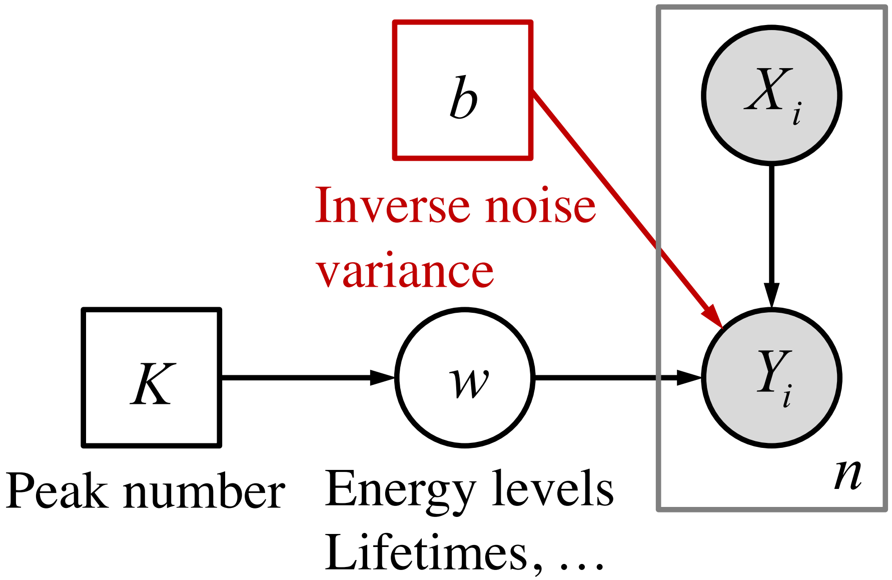
Posterior distribution

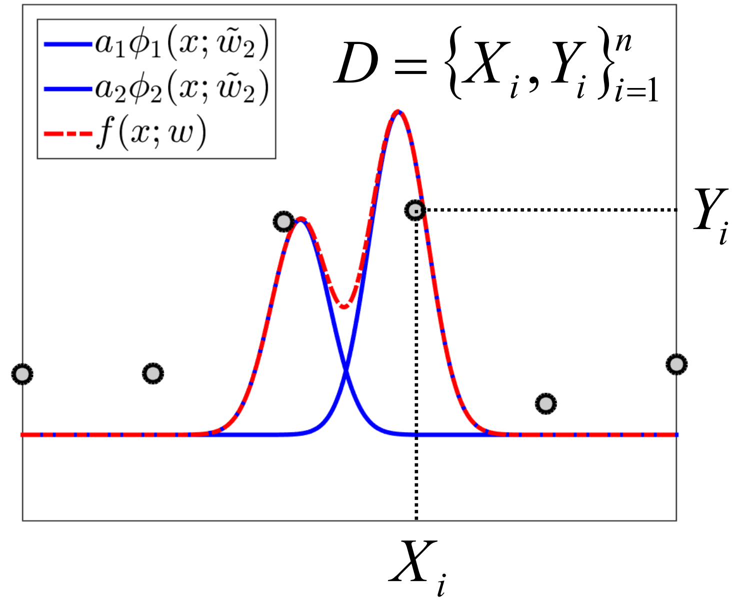


$$p(w|D,K,b)=\frac{p(Y^n|X^n,w,b)p(w|K)}{p(Y^n|X^n,K,b)}=\frac{1}{\tilde{Z}_n(K,b)}\exp(-nbE_n(w))\varphi(w|K)$$

What is the optimal pair of K and b?

Bayesian hierarchical modeling





[1] Regression

$$p(w|D,K,b)=\frac{p(Y^n|X^n,w,b)p(w|K)}{p(Y^n|X^n,K,b)}$$

$$p(Y^n|X^n,w,b)=\left(\frac{b}{2\pi}\right)^{n/2}\exp(-nbE_n(w))$$

$$p(Y^n|X^n,K,b)=\int dw\, p(Y^n|X^n,w,b)p(w|K)$$

[2] Model Selection

$$p(K,b|D)=\frac{p(Y^n|X^n,K,b)p(K)p(b)}{p(Y^n|X^n)}$$

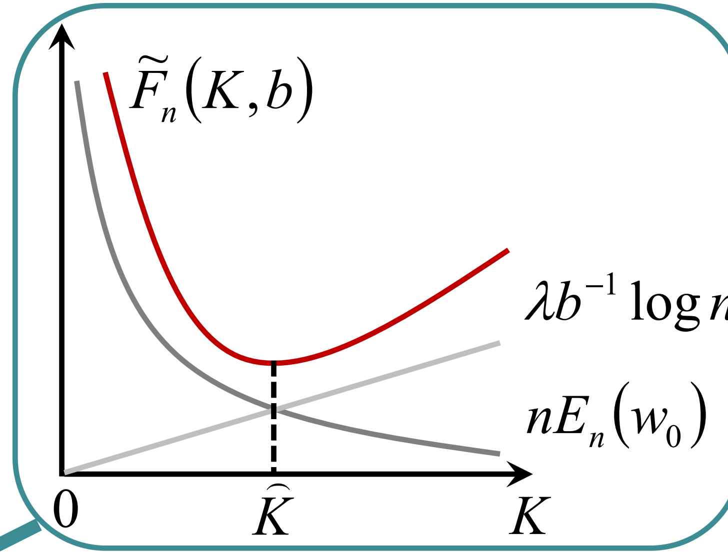
$$p(Y^n|X^n)=\sum_K\int db\, p(Y^n|X^n,K,b)p(K)p(b)$$

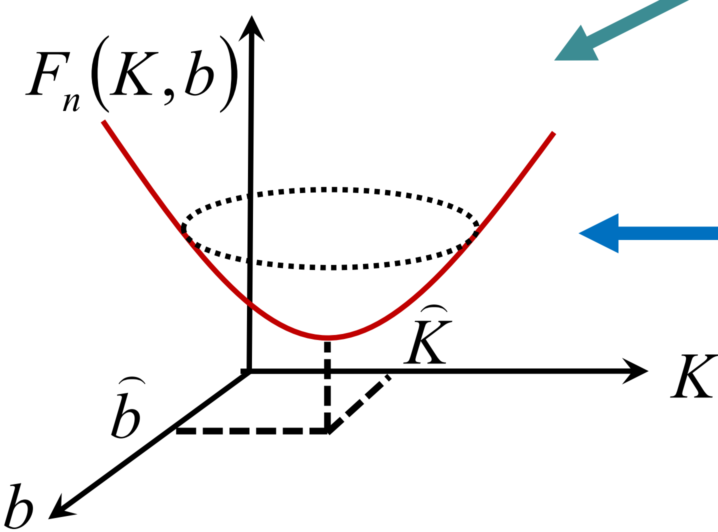
$$(\hat{K},\hat{b}):=\arg\max_{K,b} p(Y^n|X^n,K,b)$$

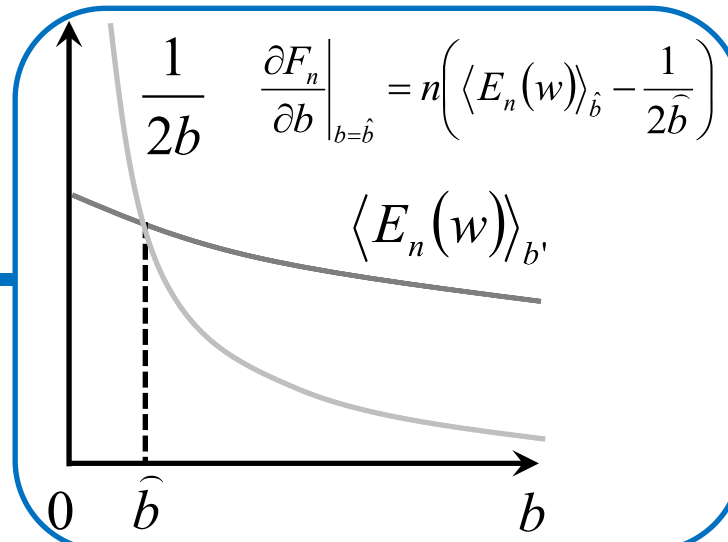
Bayes free energy

$$F_n(K,b) := -\log p\left(Y^n \middle| X^n, K, b\right)$$
$$= b\tilde{F}_n(K,b) - \frac{n}{2}\left(\log b - \log 2\pi\right)$$
$$\tilde{F}_n(K,b) := -\frac{1}{b}\log\int dw \exp\left(-nbE_n(w)\right)\varphi\left(w \middle| K\right)$$
$$= nE(w_0) + \lambda b^{-1}\log n + O_p(\log\log n)$$

[Watanabe, *Neural Compt.*, 2001]







Mathematical correspondence

Statistical physics

$$Z_N(\beta) = \int d\omega \exp(-N\beta H_N(\omega))$$

$$F_N(\beta) = -\frac{1}{\beta} \log Z_N(\beta)$$

N: Particle number

H_N: Hamiltonian

β: Inverse temperature

Bayesian inference

$$\widetilde{Z}_n(b) = \int dw \exp(-nbE_n(w)) p(w)$$

$$\widetilde{F}_n(b) = -\frac{1}{b} \log \widetilde{Z}_n(b)$$

n: Sample size

E_n: Square error

b: Inverse noise variance

Mathematical correspondence

Statistical physics

Bayesian inference

Utilize the calculation methods of free energy in statistical physics for Bayesian inference thanks to their mathematical equivalence

N: Particle number

n: Sample size

H_N: Hamiltonian

E_n: Square error

β: Inverse temperature

b: Inverse noise variance

Thermodynamic integration

$$F_n(K,b) := -\log p\left(Y^n \middle| X^n, K, b\right)$$

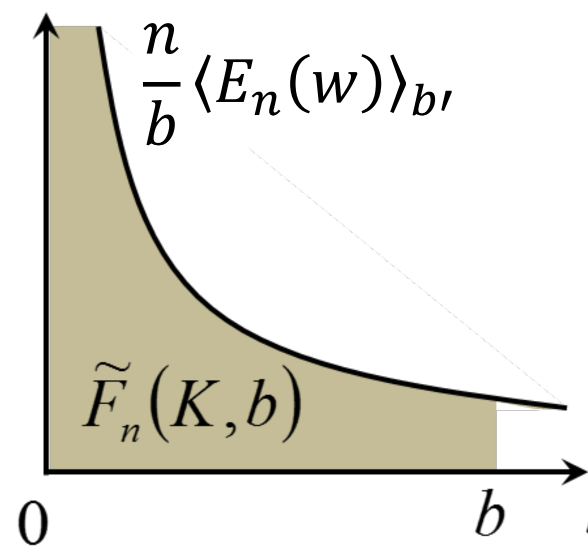
$$= b\widetilde{F}_n(K,b) - \frac{n}{2}\left(\log b - \log 2\pi\right)$$

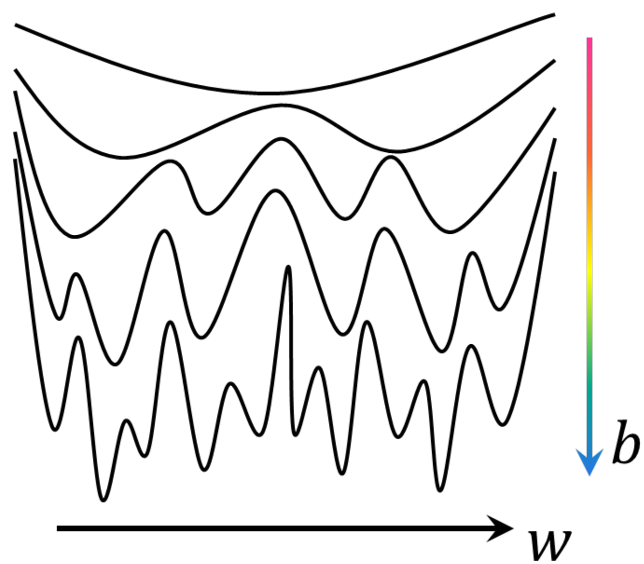
$$\widetilde{F}_n(K,b) := -\frac{1}{b}\log\int dw \exp\left(-nbE_n(w)\right)\varphi\left(w \middle| K\right)$$

$$= \frac{1}{b}\int_0^b db' \left\langle nE_n(w)\right\rangle_{b'}$$

$$\left\langle E_n(w)\right\rangle_{b'} := \int dw\, E_n(w) p\left(w \middle| D, K, b'\right)$$

$$\boxed{\begin{aligned}\left(\hat{K},\hat{b}\right) &:= \arg\max p\left(Y^n \middle| X^n, K, b\right)\\ &= \arg\min_{K,b} F_n(K,b)\end{aligned}}$$





Need to compute "thermal average"

Exchange Monte Carlo method

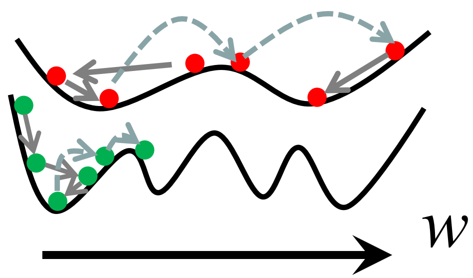
[Hukushima & Nemoto, *J. Phys. Soc. Jpn.*,1996]

$$p\left(\{w_l\}_{l=1}^{L} \middle| D, K, \{b_l\}_{l=1}^{L}\right) = \prod_{l=1}^{L} p\left(w_l \middle| D, K, b_l\right)$$

$$0 = b_1 < b_2 < \cdots < b_L = b_{\max}$$

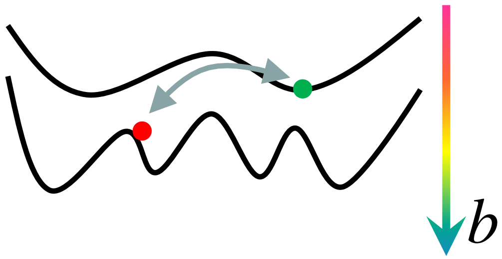
$$\boxed{\langle E_n(w)\rangle_{b_l} = \frac{1}{M_l}\sum_{t=1}^{M_l} E_n(w_l^t)}$$

1. State update (Metropolis type)

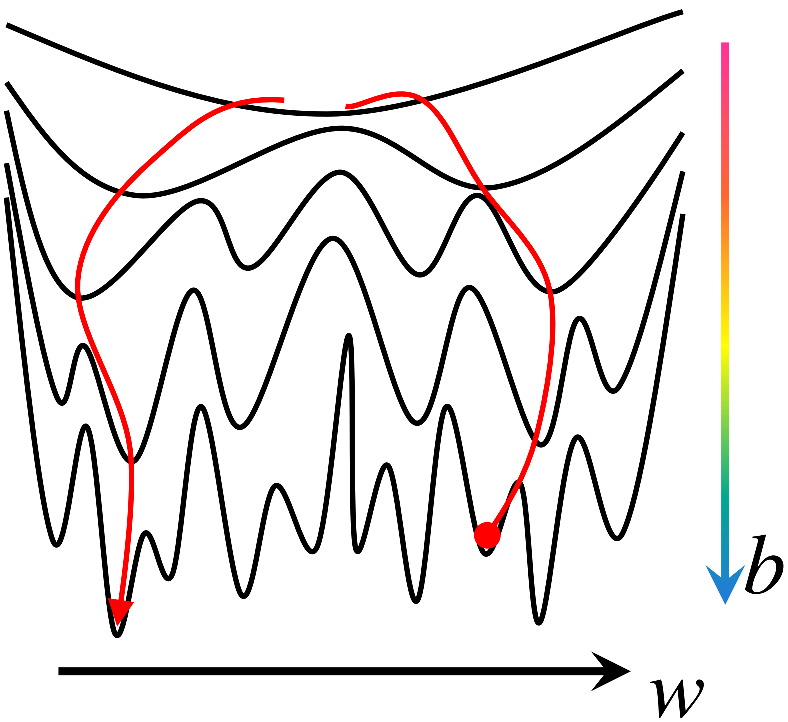


$$w_l^t \Rightarrow w_l^{t+1}$$

2. State exchange (Metropolis type)



$$\left\{w_l^{t+1}, w_{l+1}^{t+1}\right\} \Rightarrow \left\{w_{l+1}^{t+1}, w_l^{t+1}\right\}$$



Exchange Monte Carlo method

[Hukushima & Nemoto, *J. Phys. Soc. Jpn.*,1996]

$$p\left(\{w_l\}_{l=1}^{L} \middle| D, K, \{b_l\}_{l=1}^{L}\right) = \prod_{l=1}^{L} p\left(w_l \middle| D, K, b_l\right)$$

$$0 = b_1 < b_2 < \cdots < b_L = b_{\max}$$

$$\langle E_n(w) \rangle_{b_l} = \frac{1}{M_l} \sum_{t=1}^{M_l} E_n(w_l^t)$$

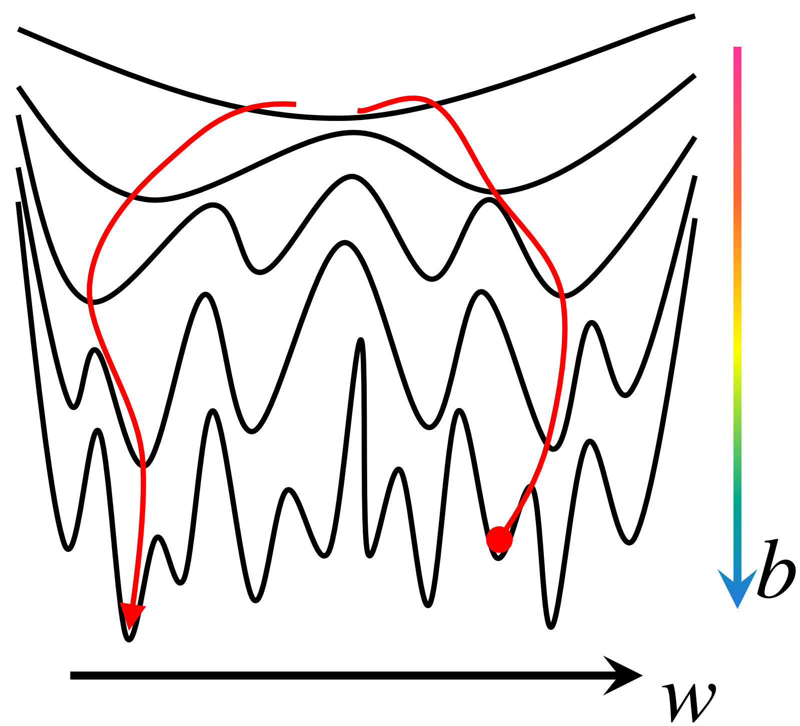
Advantages

Fast relaxation and good match for parallel computing

Disadvantages

Only a finite number L of candidates of the optimal value $b = \hat{b} \in [0, b_{max}]$, which minimize $F_n(K, b)$

How about interpolation?



Multiple histogram method

[Ferrenberg & Swendsen, *Phys. Rev. Lett.*, 1989]

$$\tilde{Z}_n(K,b) := \int dw \exp(-nbE_n(w))\varphi(w|K)$$

$$= \int dE\, g(E)\exp(-nbE)$$

Iteration

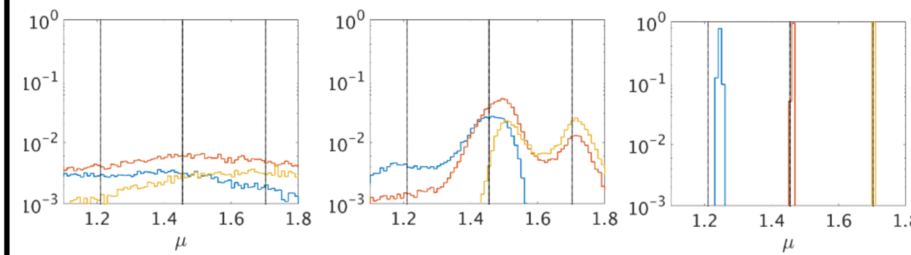
$$g(E) := \int dw\, \delta(E-E_n(w))\varphi(w|K)$$

$$= \frac{\sum_{l=1}^{L} N_l(E)}{\sum_{l'=1}^{L} M_{l'} \tilde{Z}_n(K,b_{l'})^{-1} \exp(nb_{l'}E)}$$

$N_l(E)$: Histogram of square error

M_l : Sample size of MCMC

Histogram at each temp. (Discrete)

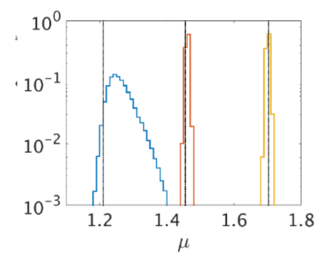


b

Small b_1 b_2 b_3 Large

Interpolation

Density at arbitrary temp. (Continuous)



b

Small $\hat{b}$ Large

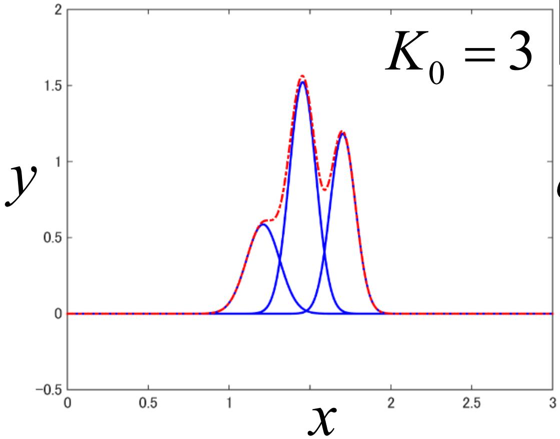
Demonstration:

Why the joint estimate of peak number and noise variance

Simulation

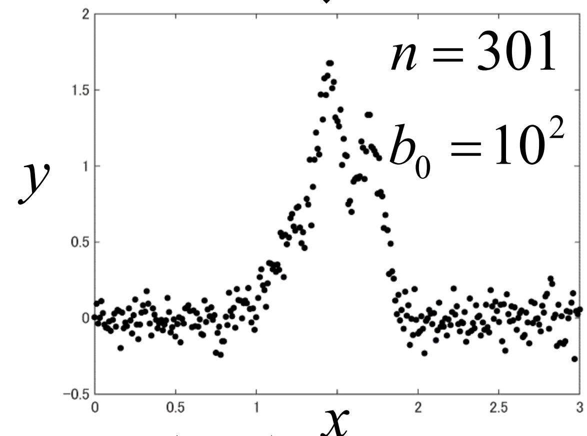
Physics

Information



Measurement (forward problem)

$$q\left(Y^n \middle| X^n, w_0, b_0\right) = \left(\frac{b_0}{2\pi}\right)^{\frac{n}{2}} \exp\left(-n b_0 E_n(w_0)\right)$$



$$\sum_{k=1}^{K_0} a_k^{*} \exp\left(-\frac{\left(x_i - \mu_k^{*}\right)^2}{2\rho_k^{*2}}\right)$$

$$w_0 := \left\{a_k^{*}, \mu_k^{*}, \rho_k^{*}\right\}_{k=1}^{K_0}$$

$$\left(\hat{K}, \hat{b}\right) = \underset{K,b}{\arg\min}\, F_n(K, b)$$

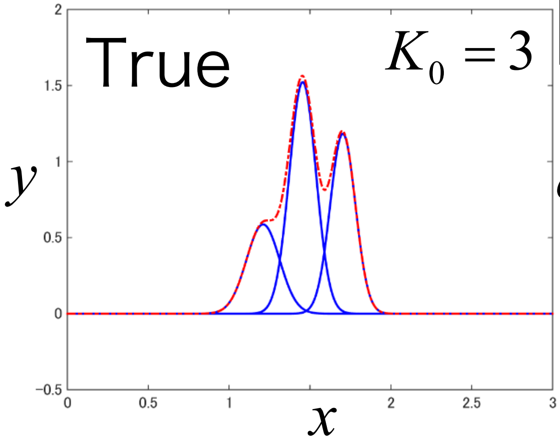
$$p\left(w \middle| D, \hat{K}, \hat{b}\right) = \frac{1}{\tilde{Z}_n} \exp\left(-n\hat{b} E_n(w)\right) \varphi\left(w \middle| \hat{K}\right)$$

Estimate (inverse problem)

Simulation

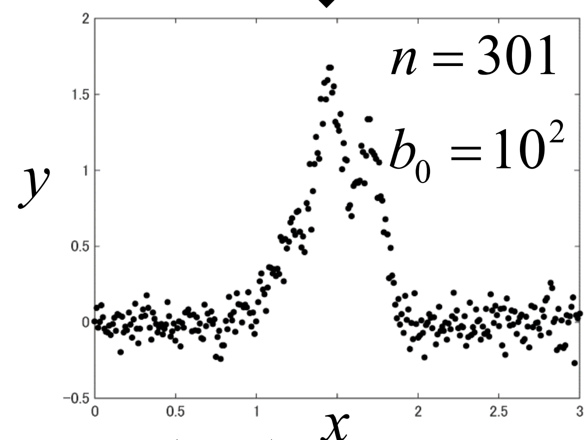
Physics

Information



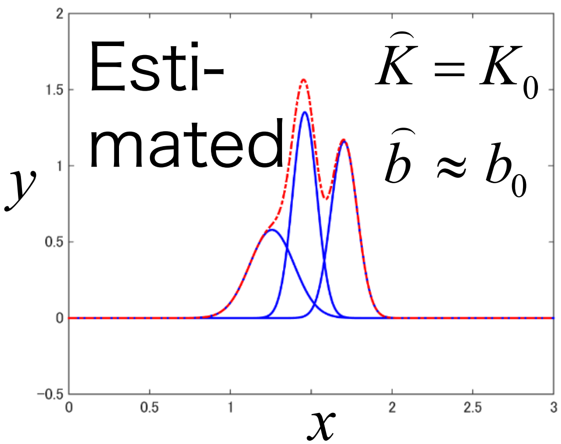
Measurement (forward problem)

$$q\left(Y^n \middle| X^n, w_0, b_0\right) = \left(\frac{b_0}{2\pi}\right)^{\frac{n}{2}} \exp\left(-n b_0 E_n(w_0)\right)$$



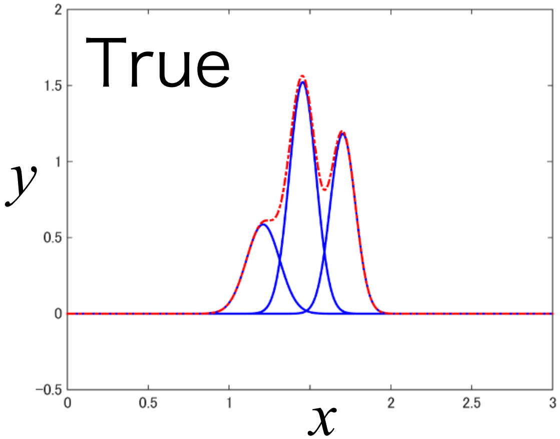
$$\left(\hat{K}, \hat{b}\right) = \arg\min_{K,b} F_n(K, b)$$

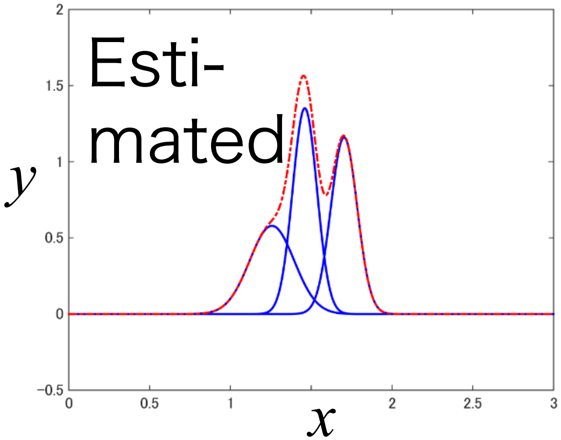
$$p\left(w \middle| D, \hat{K}, \hat{b}\right) = \frac{1}{\tilde{Z}_n} \exp\left(-n\hat{b} E_n(w)\right) \varphi\left(w \middle| \hat{K}\right)$$



Estimate (inverse problem)

Result



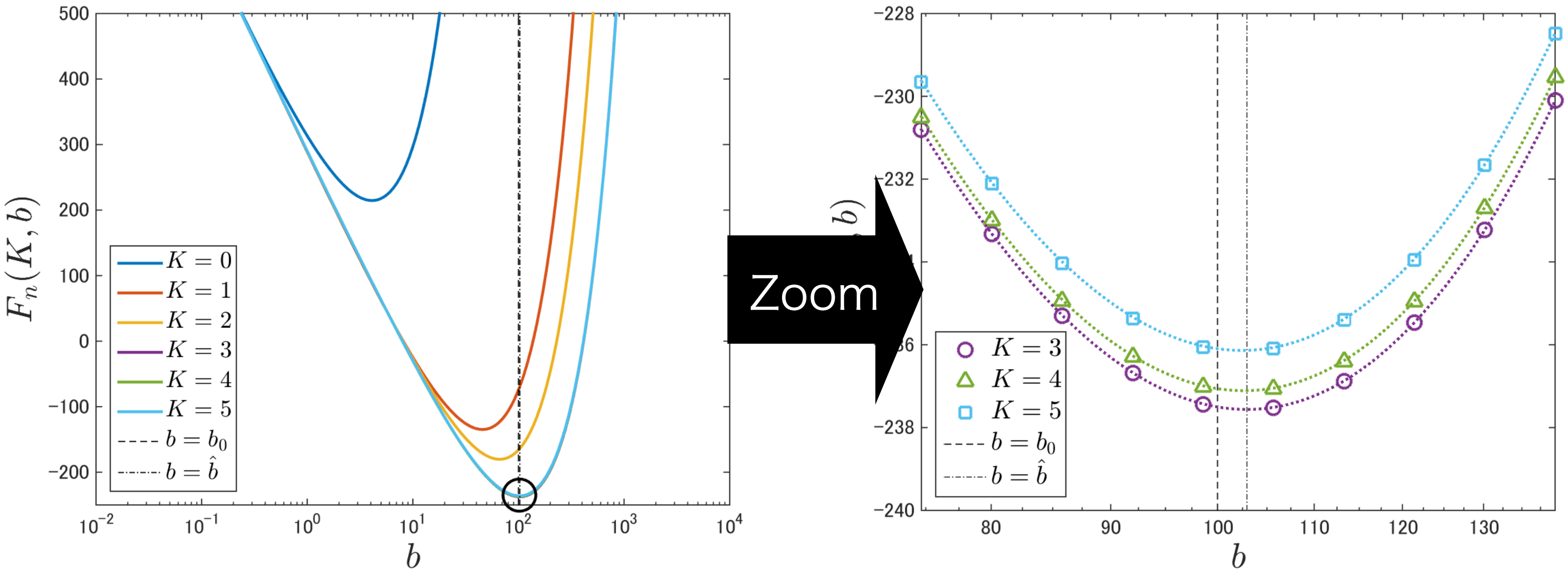


		a_k	μ_k	ρ_k
Peak 1 (*k*=1)	Est.	0.579±0.054	1.257±0.040	0.144±0.025
	True	0.587	1.210	0.102
Peak 2 (*k*=2)	Est.	1.351±0.152	1.461±0.004	0.076±0.006
	True	1.522	1.455	0.083
Peak 3 (*k*=3)	Est.	1.160±0.048	1.703±0.004	0.081±0.004
	True	1.183	1.703	0.078

	K	b
Est.	3	1.029×10^2
True	3	1.000×10^2

Estimate physical quantities with error

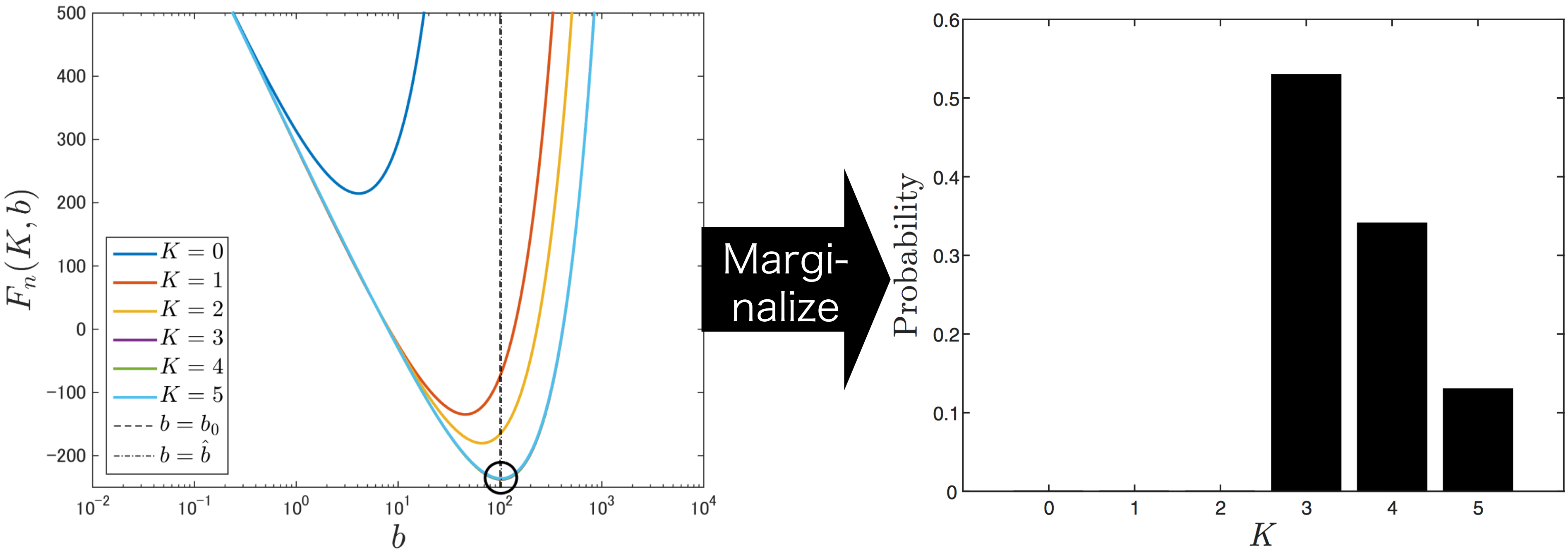
Joint estimate of peak number and noise variance



$$
\left(\hat{K},\hat{b}\right) := \arg\max p\left(Y^n \middle| X^n, K, b\right) = \mathop{\arg\min}_{K,b} F_n(K,b)
$$

Estimate correct peak number with accurate noise variance

Posterior probability of peak number

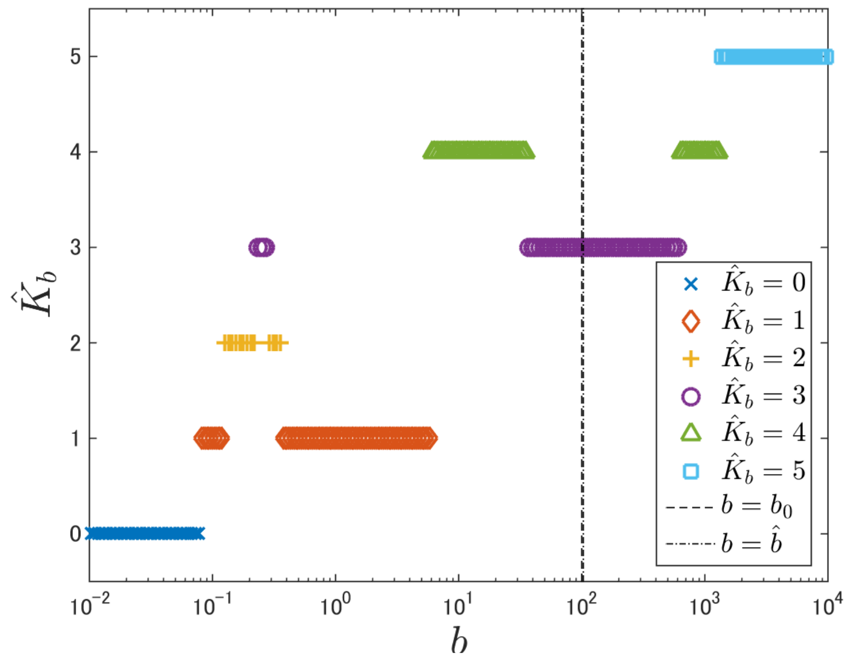


$$P(K|D) = \frac{\int_0^{b_{max}} db \exp[-F_n(K,b)]}{\sum_K \int_0^{b_{max}} db' \exp[-F_n(K,b')]}$$

$0 \le b \le b_{max}$

Discuss physical validity with these probabilities

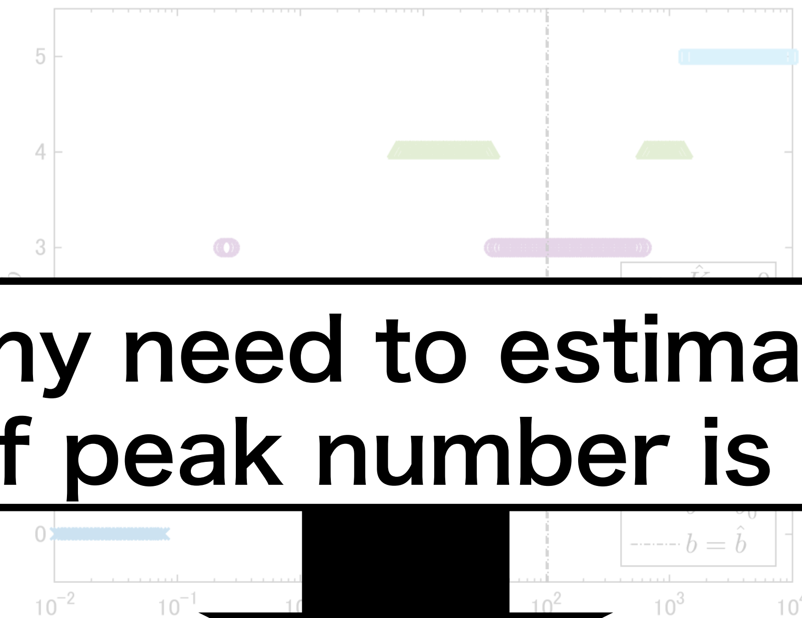
Inseparability of peak number and noise variance



Large ⟵ Noise variance ⟶ Small

$$\boxed{\begin{aligned}\widehat{K}_b &:= \arg\min_K F_n(K,b)\\ &= \arg\min_K \widetilde{F}_n(K,b)\end{aligned}}$$

Inseparability of peak number and noise variance



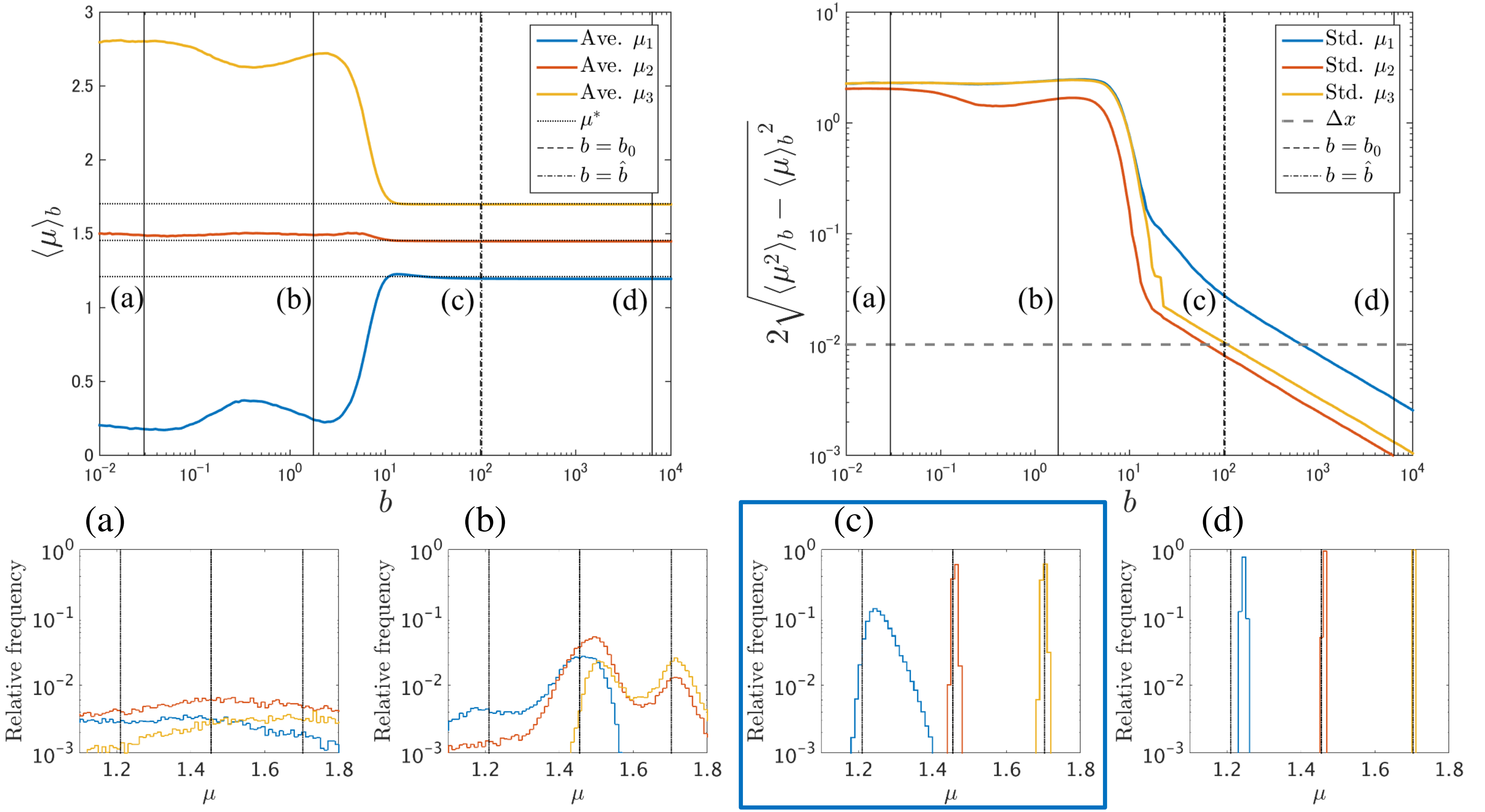
Is there any need to estimate noise variance if peak number is known?

Large ← Noise variance → Small

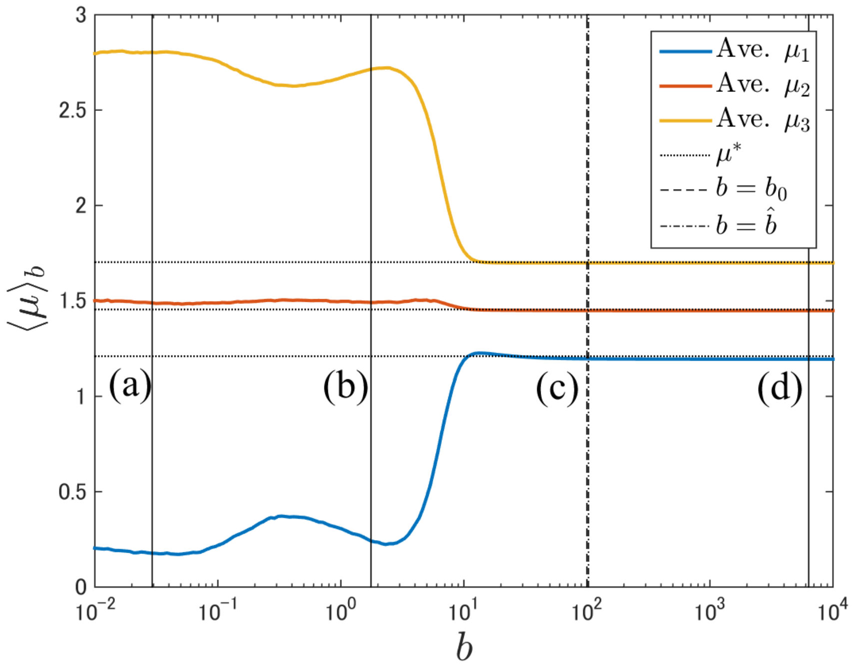
Yes!

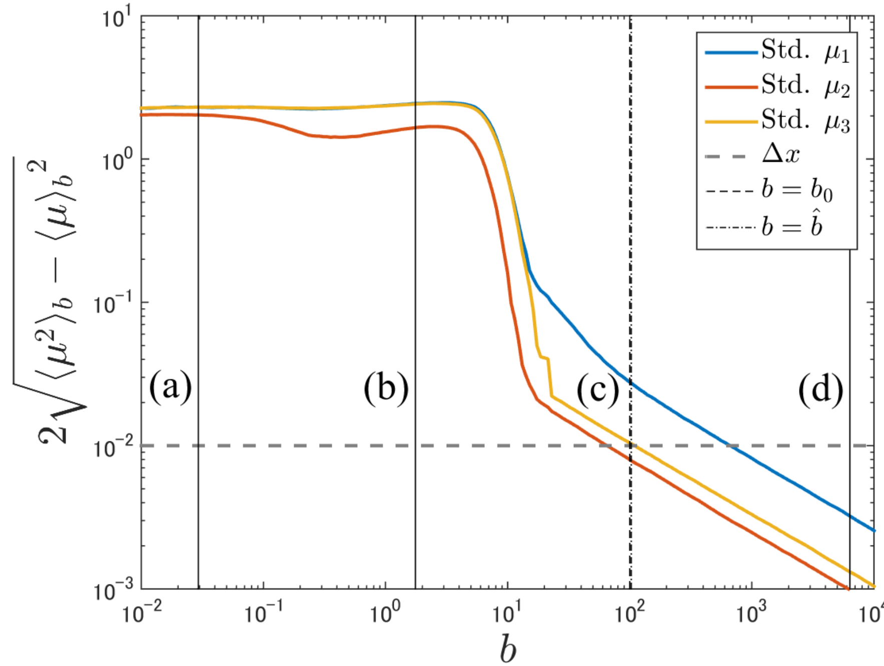
$$\widehat{K}_b := \arg\min_K F_n(K,b) = \arg\min_K \widetilde{F}_n(K,b)$$

Posterior distribution of energy levels



Posterior distribution of energy levels





Estimate of energy levels depends on the setting of noise variance even if the peak number is known

You should estimate noise variance jointly!

Summary

- Propose the framework that enables the joint estimate of peak number and noise variance from the observed spectrum by modifying the previous framework of Bayesian spectral deconvolution
  - Utilize the relationship between Bayesian inference and statistical physics
  - Show the inseparability of the estimate of peak number and noise variance

ST, K. Nagata and M. Okada, *J. Phys. Soc. Jpn*. 86, 024001, 2017. (arXiv:1607.07590)