

Variable-rate, high-rate theory: mismatch and clustering Gauss mixture models

- Review high rate, variable-rate result
- Asymptotically optimal quantization
- Worst case quantization
- Quantizer mismatch
- A universal coding result
- Robust quantization
- Lloyd clustering of Gauss mixture models

Reading: R.M. Gray and T. Linder, "Mismatch in high rate entropy constrained vector quantization," Vol. 49, pp. 1204–1217, *IEEE Trans. Inform. Theory*, May, 2003,
 A. Aiyer, K.-S. Pyun, Y.-Z. Huang, D. B. O'Brien, and R. M. Gray, "Lloyd Clustering of Gauss Mixture Models for Image Compression and Classification," *Signal Processing: Image Communication*, Vol. 20, June 2005, pp. 459–485.

All for variable-rate coding ($\eta = 0$), no comparable results exist (yet) for fixed-rate or combined constraint.

Recall high-rate, variable-rate Lagrangian result:

Theorem. Assume that f is absolutely continuous wrt Lebesgue measure, that $h(f)$ is finite, and for some Δ a partition into cubes of side Δ has finite entropy, then

$$\lim_{\lambda \rightarrow 0} \left(\inf_q \left(\underbrace{\frac{E_f[d(X, \mathcal{D}(\mathcal{E}(X)))]}{\lambda} + E_f \ell(\mathcal{E}(X))}_{\rho(f, \lambda, q) / \lambda} \right) + \frac{k}{2} \ln \lambda \right) = h(f) + \theta_k \quad (76)$$

where

$$\theta_k \triangleq \inf_{\lambda > 0} \left(\frac{\rho(u_1, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) \quad (77)$$

and u_1 is the uniform pdf on the k -dimensional unit cube

Intuitively: $\rho(f, \lambda) \approx \lambda \theta_k + \lambda h(f) - \frac{k}{2} \lambda \ln \lambda$.

Asymptotically Optimal Quantizers:

Theorem for variable-rate codes implies that that given a pdf f , there is an *asymptotically optimal* sequence of quantizers: for any decreasing sequence λ_n converging to 0 there exists a sequence of quantizers $q_n = (\mathcal{E}_n, \mathcal{D}_n, \ell_n)$ such that

$$\lim_{n \rightarrow \infty} \left(\left(\frac{E[d(X, \mathcal{D}_n(\mathcal{E}_n(X)))]}{\lambda_n} + E \ell_n(\mathcal{E}_n(X)) \right) + \frac{k}{2} \ln \lambda_n \right) = h(f) + \theta_k$$

Worst case densities

If pdf has fixed finite support Ω with volume $V(\Omega)$, then *worst case* (biggest differential entropy) is **uniform** pdf on Ω :

$$f(x) = \frac{1}{V(\Omega)}, x \in \Omega, \quad h(f) = \log V(\Omega)$$

(prove using divergence inequality)

If know mean $\mu = EX$ and covariance $K = E[(X - \mu)(X - \mu)^t]$ of the source, then *worst case* is **Gaussian** pdf:

$$f(x) = \mathcal{N}(x, \mu, K) = \frac{1}{(2\pi)^{\frac{k}{2}} |K|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^t K^{-1}(x - \mu)\right)$$

$$h(f) = \frac{1}{2} \ln(2\pi e)^k |K|,$$

High rate analog to Sakrison's Shannon rate-distortion result

What if we design asymptotically optimal sequence of quantizers q_n for pdf g (e.g., worst case), but apply it to f ? \Rightarrow **mismatch**

Mismatch

Optimize a code for a distribution P_g on \mathfrak{R}^k , but apply the code to a distribution P_f

Classic example:

- Lossless source code
- Distributions are discrete, described by pmfs g and f .

Uniquely decodable lossless code must have a collection of codeword lengths $\ell(i)$ in nats that satisfies the Kraft inequality

If a discrete source has pmf $g = \{g_i\}$ with Shannon entropy

$$H(g) = -\sum_i g_i \ln g_i,$$

divergence inequality \Rightarrow If ℓ admissible,

$$E_g \ell = \sum_i g_i \ell(i) \geq H(g), \text{ with equality if}$$

$$\ell(i) = -\ln p_i \text{ (Ignore constraint of integer lengths)}$$

Apply optimal code for pmf g instead to pmf f :

$$\begin{aligned} E_f \ell &= \sum_i \ell(i) f_i = -\sum_i f_i \ln g_i \\ &= H(f) + \sum_i f_i \ln \frac{f_i}{g_i} \triangleq H(f) + H(f||g), \end{aligned}$$

Extend mismatch idea to fixed dimension high rate vector quantization.

For fixed rate codes, done by Bucklew (1984)

Theorem. Suppose that q_n is asymptotically optimal for a pdf g and that f is a pdf satisfying an horrendously complicated condition given by Bucklew, the only simple version of which is that f/g is bounded. Then

$$\lim_{R \rightarrow \infty} 2^{R/k} D_f(q_n) = a_k \left(\int \frac{f(x)}{g(x)^{2/2+k}} dx \right) \left(\int g(x)^{k/k+2} dx \right)^{2/k}$$

Intuition: This is the result that follows from Gersho's conjecture using approximations to integrals and a quantizer point density.

In fact, as part of the proof, Bucklew demonstrated the existence of a quantizer point density for fixed rate quantizers.

Can derive heuristically using Gersho's approximations: Assume $\Lambda(x)$ optimal point density function for variable-rate quantizer of g .

Optimize q for Gaussian g : Gersho's approximations imply that optimal point density function for purely variable rate coding is $\Lambda(x) = \text{constant}$ (regardless of pdf) \Rightarrow

$$D_g(q) \approx c_k \int (N(q)\Lambda(x))^{-\frac{2}{k}} f(x) dx$$

optimal N satisfies

$$N^{2/k} = \frac{2}{k\lambda} c_k E_g \left(\Lambda(X)^{-2/k} \right) \approx \left(\frac{2}{k\lambda} D_g(q) \right)$$

so that

$$D_g(q) \approx \frac{k\lambda}{2}$$

Variable rate case:

Theorem. *The mismatch theorem:* Suppose that q_n is asymptotically optimal for $\lambda_n \rightarrow 0$ for a pdf g and that f is pdf for which f/g is bounded, then

$$\lim_{n \rightarrow \infty} \frac{D_f(q_n)}{\lambda_n} + E_f \ell_n(\mathcal{E}_n(X)) + \frac{k}{2} \ln \lambda_n = \theta_k - \int dx f(x) \ln g(x)$$

or, equivalently,

$$\lim_{n \rightarrow \infty} \frac{D_f(q_n)}{\lambda_n} + E_f \ell_n(\mathcal{E}_n(X)) + \frac{k}{2} \ln \lambda_n = \underbrace{\theta_k + h(f)}_{\text{optimum for } f} + \underbrace{H(f||g)}_{\text{mismatch}},$$

for small λ . Note this is independent of pdf, hence $D_g(q) \approx D_f(q)$. (*robust* (Sakrison, Lapidoth))

Entropy approximation ($E_g \ell$)

$$H_g(q(X)) \approx \ln N - H(g||\Lambda) \approx \ln \left(\frac{2}{k\lambda} D_g(q) \right) - H(g||\Lambda) \approx \ln \lambda - H(g||\Lambda)$$

When apply $\Lambda(X)$ to f and use optimal length function for g ($\ell(i) = -\ln P_g(S_i)$), the average rate will be

$$\begin{aligned} - \sum_i P_f(S_i) \ln P_g(S_i) &= - \sum_i P_f(S_i) \ln P_f(S_i) + \sum_i P_f(S_i) \ln \frac{P_f(S_i)}{P_g(S_i)} \\ &\approx H_f(q(X)) + H(f||g) \end{aligned}$$

so that

$$\frac{D_f(q)}{\lambda} + E_f \ell(\mathcal{E}(X)) \approx \frac{D_g(q)}{\lambda} + H_f(q) + H(f||g)$$

Heuristically: Λ causes uniform distribution of codewords, so average distortion the same. The mismatch is all in the rate, i.e., the lossless coding.

So truly continuous generalization of classic lossless coding mismatch.

The mismatch theorem implies that if q_n asymptotically optimal for g , then when applied to f it will yield *the asymptotically optimal performance for f plus $H(f||g)$* .

Suppose f and g have means μ_f and μ_g and covariances K_f and K_g , and g alone is assumed to be Gaussian. Then

$$H(f||g) = h(g) - h(f) - (k/2) + (1/2) \text{Tr } K_g^{-1} K_f + (\mu_f - \mu_g)^t K_g^{-1} (\mu_f - \mu_g).$$

Thus if choose equal means and covariances, $\mu_g = \mu_f = \mu$ and $K_f = K_g = K$, then

$$h(f) - h(g) + H(f||g) = 0$$

or $H(f||g) = h(g) - h(f)$

High Rate Variable Rate Universal Coding

An aside – a “universal coding” result. Makes part of Gersho heuristic rigorous.

Corollary 2. *Suppose that $q_n = (\mathcal{E}_n, \mathcal{D}_n, \ell_n)$ is a sequence of variable rate quantizers that is asymptotically optimal for a pdf g for some decreasing sequence $\lambda_n \rightarrow 0$. Assume also that f is a pdf that meets the condition of the mismatch theorem. Define ℓ'_n to be the optimal length function for \mathcal{E}_n and P_f . Then $q'_n = (\mathcal{E}_n, \mathcal{D}_n, \ell'_n)$ is asymptotically optimal for P_f , i.e., $\lim_{n \rightarrow \infty} \theta(f, \lambda_n, q'_n) = \theta_k$.*

The length function of the quantizer matched to the true source, but encoder not optimized for the new length function. Thus there remains a mismatch in the code sequence, which nonetheless is asymptotically optimal!

Mismatch as Distortion between pdfs

The relative entropy quantifies the high rate mismatch from optimal performance of a quantizer optimized for a “model” pdf and then applied to a “true”

\Rightarrow adds motivation for $H(f||g)$ as a “distance” or “distortion measure” on pdfs in order to “quantize” the space of pdf’s to fit models to observed data.

Composite quantizers

A single Gaussian source provides both a worst case and a robust design for a source whose second order properties are known.

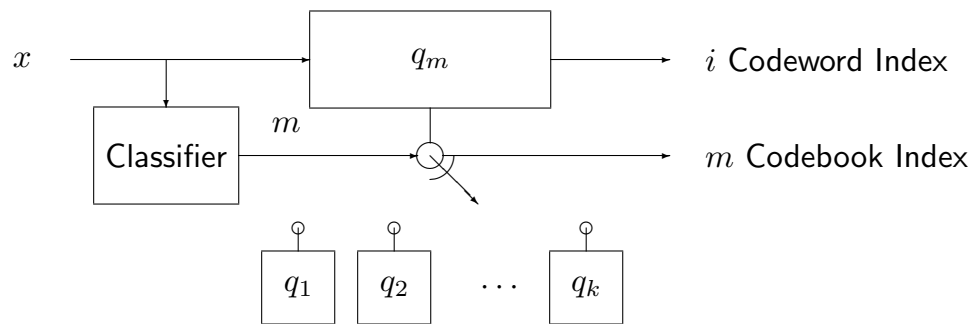
Too conservative.

Speech tradition is partition input space, fit separate Gaussian models to different cells. Use Gaussian code i for signals in cell i . Gives Lloyd design algorithm for Gauss models.

f be the "true" pdf on Euclidean space \mathbb{R}^k .

Consider a partition $\mathcal{S} = \{S_m; m \in \mathcal{J}\}$ of \mathbb{R}^k , where $\mathcal{J} = \{1, \dots, M\}$.

Assume $P_f(S_m) > 0$ for all m .



Decoder: $\mathcal{D}_m(i)$

Design a separate code, say q_m , for each f_m , use a two-step coding procedure.

If $x \in S_m$, use q_m to quantize.

Assume collection of model pdf's $\{g_m\}$ on \mathbb{R}^k .

Given partition, consider f to be a mixture source:

$$f(x) = \sum_m p_m f_m(x),$$

where $f_m(x) = f(x)/p_m$ for $x \in S_m$ and 0 otherwise, and $p_m = P_f(S_m)$.

Composite/classified/switched quantizer:

Each g_m is a design model for f_m . Will choose to be Gaussian $g_m = \mathcal{N}(\mu_m, K_m)$ (robust worst case). Optimize $q_m = (\mathcal{E}_m, \mathcal{D}_m, \ell_m)$ for g_m .

Assume λ small (high-rate regime)

$$\rho(\lambda, g_m, q_m) \approx \rho(\lambda, g_m)$$

Overall quantizer: Two-step (classified) quantizer:

If input vector $x \in S_m$, quantize x using q_m .

Encoder sends codeword index along with codebook index m to decoder.

$$\mathcal{E}(x) = (m, \mathcal{E}_m(x)) \text{ if } x \in S_m$$

$$\mathcal{D}(m, i) = \mathcal{D}_m(i)$$

$$\ell(m, i) = L(m) + \ell_m(i)$$

where L a length function (satisfies Kraft) for index indicating which cell (which quantizer chosen). Can show optimal choice is $L(m) = -\ln P_X(S_i)$

Straightforward to show:

$$\frac{\rho(f, \lambda, q)}{\lambda} - h(f) = \sum_m p_m \left(\frac{E_{f_m} d(X, \mathcal{D}(\mathcal{E}(X)))}{\lambda} + E_{f_m} \ell(\mathcal{E}(X)) - h(f_m) \right)$$

where Z has distribution $\Pr(Z = m) = P_X(S_i) = p_m$.

With the optimal choice of $L(m) = -\ln p_m$, the average code

which =0 if Gaussian moments match those of f_m !

Assume matched moments, mismatch theorem \Rightarrow

$$\rho(f, \lambda, Q)/\lambda - \rho(f, \lambda)/\lambda \approx \sum_m p_m H(f_m || g_m). \tag{78}$$

Want to choose \mathcal{S} and models g_m to minimize.

length of the composite quantizer is

$$EL(Z) + \sum_m p_m E_f \ell_m(\mathcal{E}_m(X)) = H(Z) + \sum_m p_m E_f \ell_m(\mathcal{E}_m(X))$$

With this choice

$$\rho(f, \lambda, q)/\lambda - h(f) = \sum_m p_m (\rho(f_m, \lambda, q_m)/\lambda - h(f_m)) .$$

Since q_m is optimized for Gaussian model g_m and applied to the pdf f_m , mismatch theorem \Rightarrow

$$\begin{aligned} \rho(f_m, \lambda, q_m)/\lambda - \rho(g_m, \lambda, q_m)/\lambda &\approx h(f_m) - h(g_m) + H(f_m || g_m) \\ &= -\frac{k}{2} + \frac{1}{2} \text{Tr} K_{g_m}^{-1} K_{f_m} + (\mu_{f_m} - \mu_{g_m})^t K_{g_m}^{-1} (\mu_{f_m} - \mu_{g_m}) \end{aligned}$$

Quantizer Mismatch Distortion

Seek a collection of Gaussian pdfs $\mathcal{G} = \{g_m\}$ and a partition $\mathcal{S} = \{S_m\}$ of \mathbb{R}^k which minimizes overall mismatch $\bar{I}_f = \inf_{\mathcal{S}, \mathcal{G}} \bar{I}_f(\mathcal{S}, \mathcal{G})$, where $\bar{I}_f(\mathcal{S}, \mathcal{G}) = \sum_m P_f(S_m) H(f_m || g_m)$.

Can solve Lloyd clustering: pose as a quantization problem with encoder $a : \mathbb{R}^k \rightarrow \mathcal{J}$ described by the partition $\mathcal{S} = \{S_m\}$ by $a(x) = m$ if $x \in S_m$, $m \in \mathcal{J}$, decoder $b : \mathcal{J} \rightarrow \mathcal{M}$ defined by $b(m) = g_m$.

Decoder Given encoder index m corresponding to encoder cell S_m , best possible g_m (minimizing the mismatch) is the Gaussian solution to $g_m = \arg \min_{g \in \mathcal{M}} H(f_m || g)$, \Rightarrow Gaussian source with the same mean and covariance as f_m . With this decoder the minimum mismatch

problem becomes

$$\bar{I}_f = \inf_{\mathcal{S}} \sum_m P_f(S_m) \min_{g \in \mathcal{M}} H(f_m || g).$$

Encoder Consider distortion measure:

$$d_H(x, m) = \ln(f(x)/g_m(x)) + L(m),$$

where L satisfies the Kraft inequality.

d_I is not nonnegative, but its average with respect to f is nonnegative from the divergence inequality, meets requirements of Lloyd.

Optimal encoder is a minimum distortion encoder: $\mathcal{G} a(x) = \arg \min_m d_H(x, m) = \arg \min_m (L(m) - \ln g_m(x))$.

When using individual Gaussian models with optimal codebooks and length functions,

$$d_H(x, m) = \ln \frac{f(x)}{g_m(x)} - \ln p_m = \ln f(x) - \ln p_m + \frac{1}{2} \ln ((2\pi)^k |K_m|) + \frac{1}{2} (x - \mu_m)^t K_m^{-1} (x - \mu_m)$$

where μ_m and K_m are the mean and covariance of the Gaussian pdf g_m .

$\ln f(x)$ term and constant terms have no effect on encoder a

Corresponding encoder partition \mathcal{S} yields average distortion

$$\int dx f(x) d_I(x, a(x)) = \sum_m p_m \left(L(m) + \int_{S_m} dx f_m(x) \ln \frac{f_m(x) p_m}{g_m(x)} \right)$$

Thus

$$\begin{aligned} \int dx f(x) d_I(x, a(x)) &= \sum_m p_m H(f_m || g_m) + \sum_m p_m \ln \frac{p_m}{e^{-L(m)}} \\ &\geq \sum_m p_m H(f_m || g_m) \end{aligned}$$

with equality iff choose $L(m) = -\ln p_m$.

Average distortion according to d_I is exactly the mismatch which we wish minimize.

Define the *quantizer mismatch* distortion by

$$\begin{aligned} d_{QM}(x, m) &= \ln \frac{1}{g_m(x)} - \ln p_m - \frac{k}{2} \ln(2\pi) \\ &= \frac{1}{2} \ln |K_m| + \frac{1}{2} (x - \mu_m)^t K_m^{-1} (x - \mu_m) - \ln p_m, \end{aligned}$$

Similar distortion measures arise in other contexts: “maximum likelihood” or “log likelihood,” minimum discrimination distortion, Itakura-Saito distortion in speech have this form.

Variation: change weighting:

$$d_{QM,\lambda}(x, m) = \frac{1}{2} \ln |K_m| + \frac{1}{2} (x - \mu_m)^t K_m^{-1} (x - \mu_m) - \lambda \ln p_m. \quad (79)$$

The added flexibility allows improved control over the number of components in the Gauss mixture.

Toy example: Gauss Mixture Design

Can use Lloyd algorithm to optimize. Outcome of optimization is a *Gauss mixture*.

Alternative to Baum-Welch/EM algorithm.

Lloyd conditions applied to quantizer mismatch distortion:

Encoder (partition) $\alpha(x) = \operatorname{argmin}_m d_{\text{QM}}(x, m)$

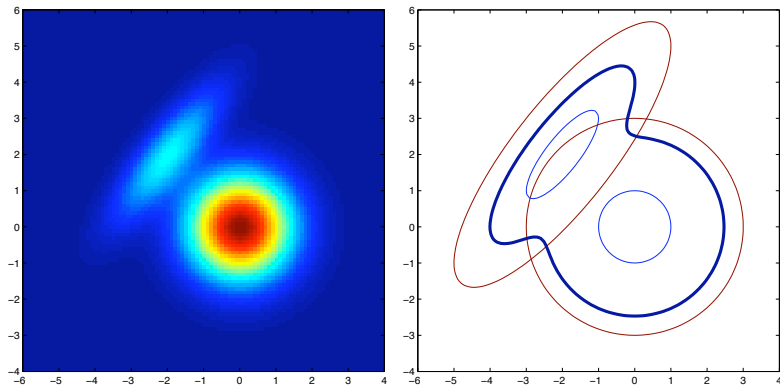
Decoder (centroid) $g_m = \operatorname{argmin}_{g \in \mathcal{M}} H(f_m || g) = \mathcal{N}(\mu_m, K_m)$.

Length Function $L(m) = -\ln p_m$.

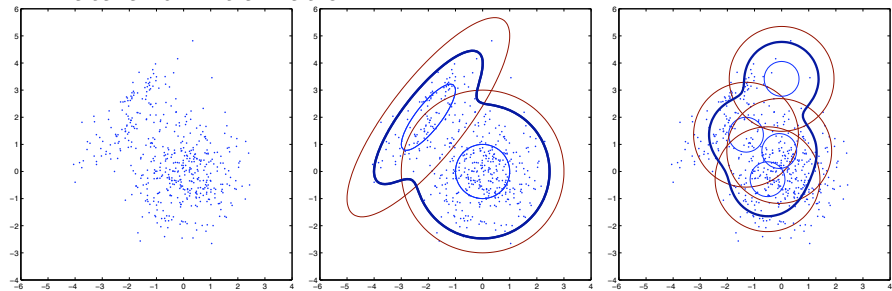
Gauss mixture vector quantization (GMVQ)

$$m_1 = (0, 0)^t, K_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, p_1 = 0.8$$

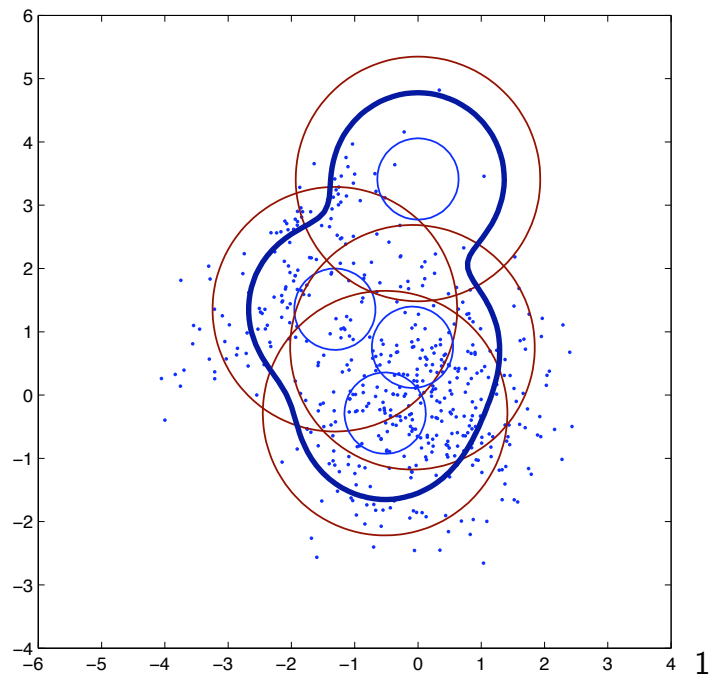
$$m_2 = (-2, 2)^t, K_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1.5 \end{bmatrix}, p_2 = 0.2$$

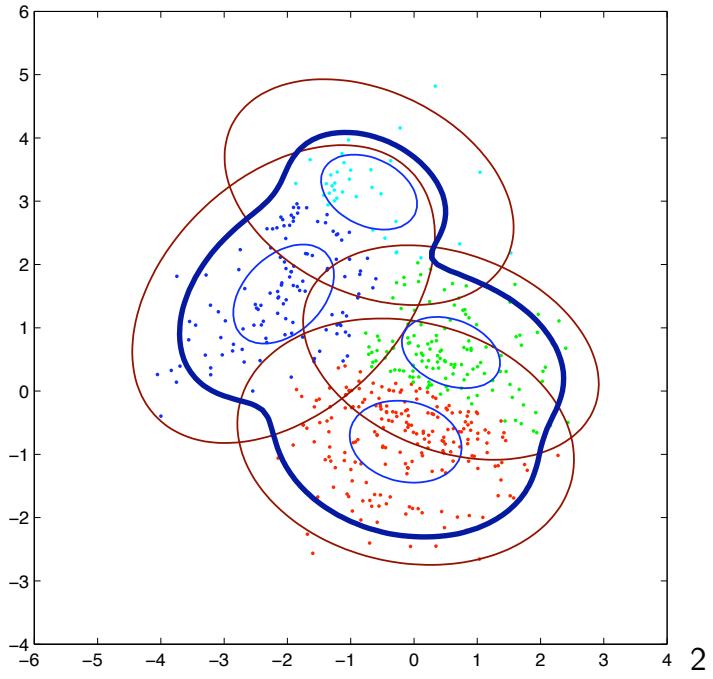


Data and Initialization:



Data True pdf superimposed on data Initial code

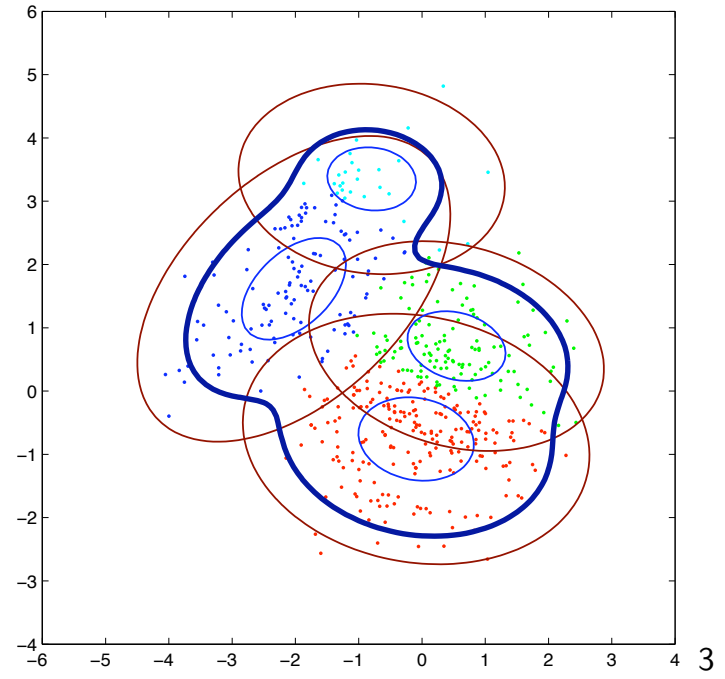




2

Quantization

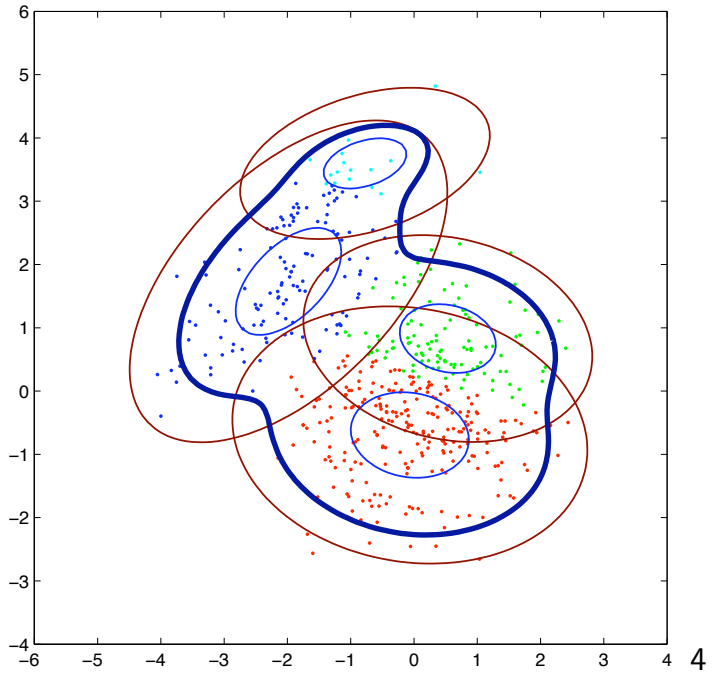
204



3

Quantization

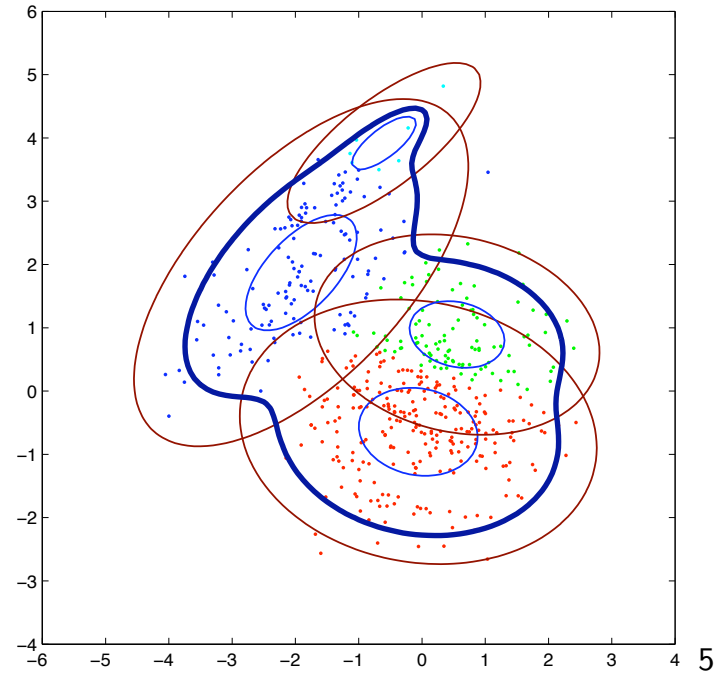
205



4

Quantization

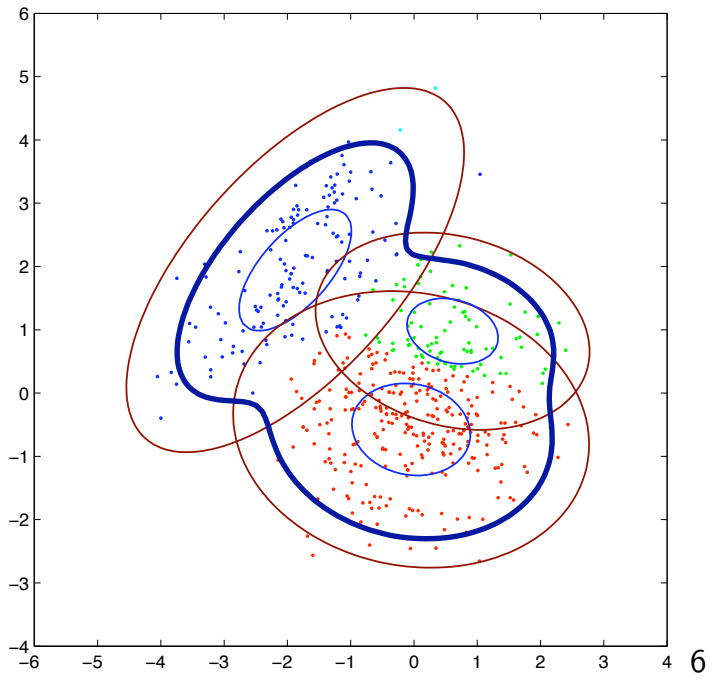
206



5

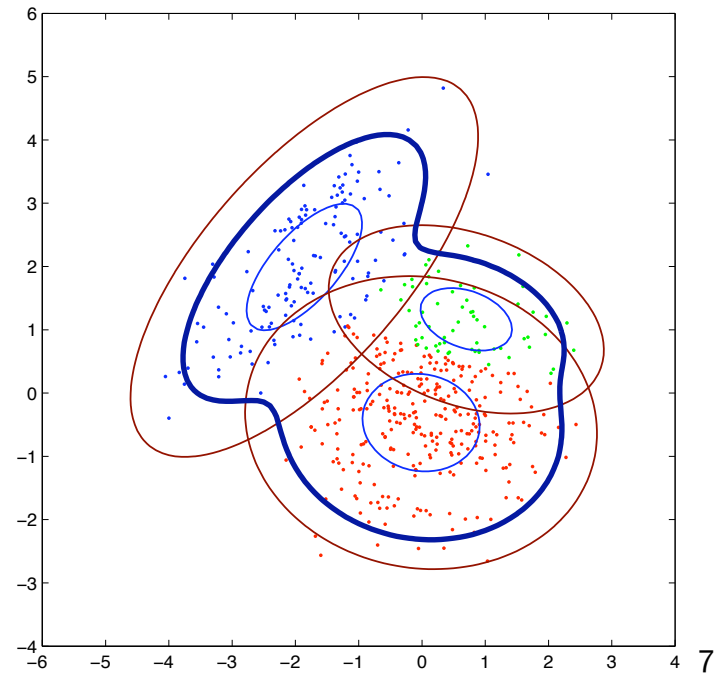
Quantization

207



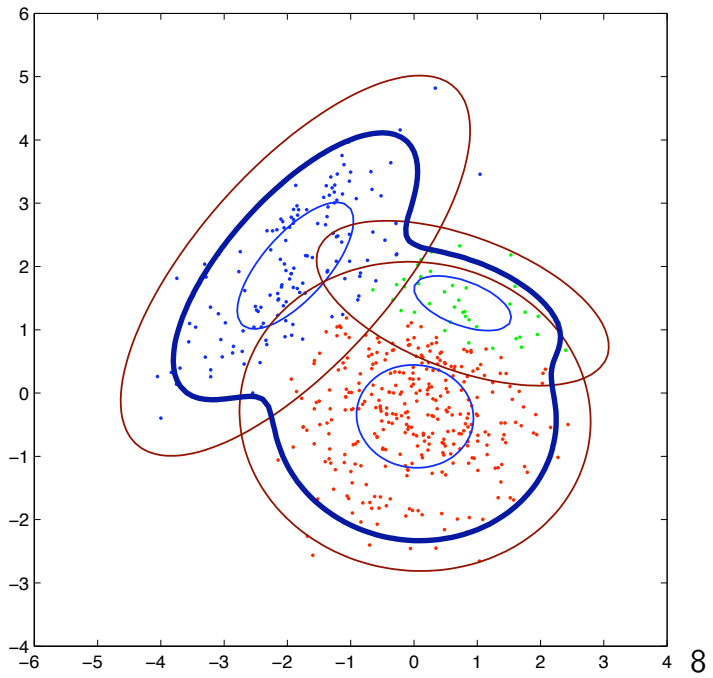
Quantization

208



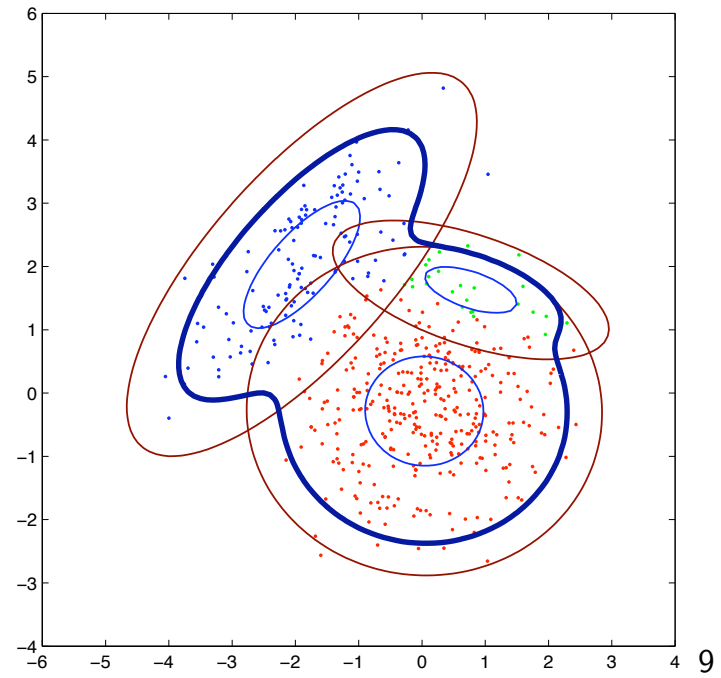
Quantization

209



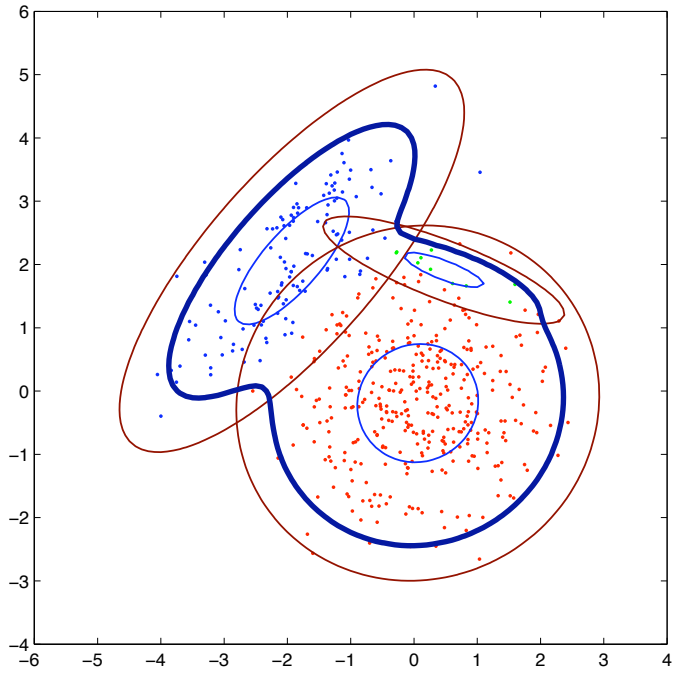
Quantization

210



Quantization

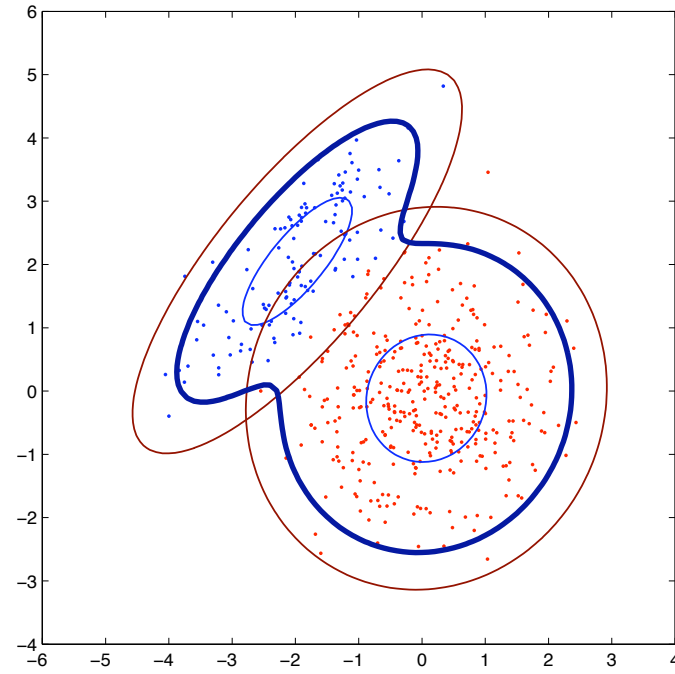
211



10

Quantization

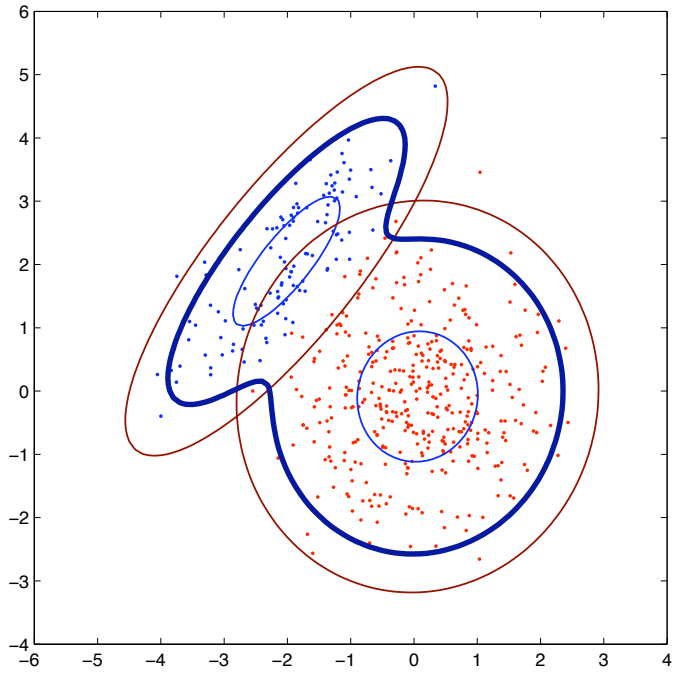
212



11

Quantization

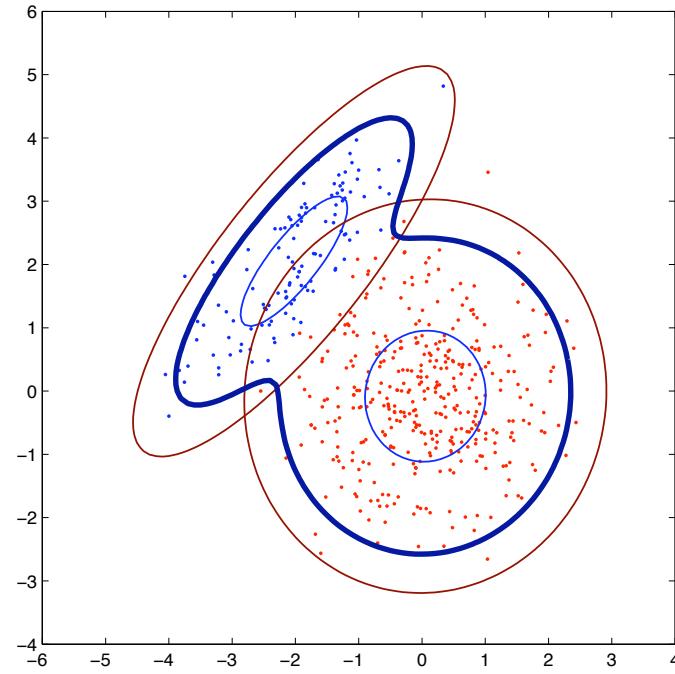
213



12

Quantization

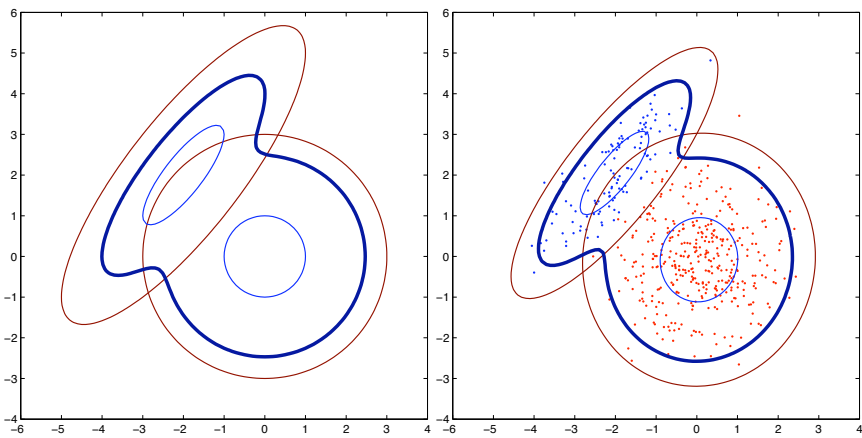
214



Converged!

Quantization

215



True source

Data driven model

Algorithm for fitting a Gauss mixture to a dataset provides a supervised learning method for classifier design: Given a collection of classes of interest, design for each a GMVQ

Given a new vector (e.g., image):

Traditional approach: plug in density estimate to optimal Bayes

Best codebook approach: Code the image using each of the Gauss mixture VQs and select the class corresponding to the quantizer with the smallest average mismatch distortion.

[Codebook-matching, classification by compression.](#)

minimum distortion or “nearest-neighbor” classifier, statistical classification without explicit estimation of $P_{Y|X}$