Bayesian learning for weakly supervised object classification

Peter Carbonetto, Gyuri Dorkó and Cordelia Schmid
INRIA Rhône-Alpes, Grenoble, France

August 5, 2004

Abstract

We explore the extent to which we can exploit interest point detectors for representing and recognising classes of objects. Detectors propose sparse sets of candidate regions based on local salience and stability criteria. However, local selection does not take into account discrimination reliability across instances in the same object class, so we realise selection by learning from weakly supervised data in the form of images paired with their captions. Through experiments on a wide variety of object classes and detectors, we show that modeling object recognition as a constrained data association problem and learning the Bayesian way by integrating over multiple hypotheses leads to sparse classifiers that outperform contemporary methods. Moreover, our learned representations based on local features leave little room for improvement on standard image databases, so we propose new data sets to corroborate models for general object recognition.

1 Introduction

The proliferation of methods for extracting and describing salient and repeatable features [14, 16, 21, 22, 24] combined with recent advances in machine learning has fostered new and robust representations of object classes [1, 9, 29]. It is widely appreciated that local features are generally inadequate for identifying and locating objects in scenes (a person is not just an elbow!), and there has been some success in learning principled representations of relations between parts [11] and global context [5, 28]. However, we believe it is still of general interest to explore the extent to which independent parts can function as a basis for robust object recognition. In this paper, we show that significant improvements on state-of-the-art local feature recognition systems can be achieved by adopting principled probabilistic models and semi-supervised Bayesian learning techniques for automatically selecting and combining features from multiple local detectors. By selecting sparse sets of discriminative scale-invariant features, our representations accurately recognise objects found in a wide variety of scenes at different poses and scales (see Fig. 1 for some examples). What's more, we learn with very little supervision from the user.

The first step in our object recognition approach is to obtain a set of local *a priori* salient regions of the scene. The original motivation for local interest regions was to provide a stable basis for recognition of objects at varying viewpoints. Local features bring tolerance to clutter, occlusions and deformations. Some of the best examples can be found in [12], since the authors demonstrate the application of local features to object matching at widely disparate locations, even in the presence of clutter. Good detectors extract a sparse set of interest regions without relinquishing information content of the scene, and select the same regions when observed at different viewpoints and scales.

In current literature, there are many definitions as to what constitutes a good scale-covariant region, predicated on maximising disparate criteria. For example, the Harris-Laplace detector [24] selects image regions that are locally precise to translation (corners). Kadir and Brady [14] hypothesize that salient regions

Figure 1: Example car images. Note that there is no image that contains only cars, and that car size varies markedly.

correspond to local maximums of complexity (or entropy). Interest point detectors were originally designed for object matching and wide baseline stereo, not for detecting semantic categories of objects. However, recent work has shown that interest regions work well for general recognition [9, 11, 29]. In this paper, we compare the ability of detectors to propose good features for recognition. We also expect that using multiple detectors will provide complementary information, hence improve recognition, a hypothesis we examine experimentally.

Interest point detectors extract regions based on local information, which is not enough to determine whether or not they are genuinely useful for recognising classes of objects. We need to learn which features are most discriminative based on some form of supervision from the user. Complete supervision requires the user to label individual features by segmenting the object from the background. Not only is this a time-consuming and equivocal task, since people tend to segment scenes differently, it also inhibits us from exploiting large quantities of captioned images available on the Internet (in the form of news photos, for example [27]). Rather, we request image labels which indicate the presence or absence of an object in each scene. For recognition of cars, the training data consists of positively labeled images containing cars, and negative images without cars (see Fig. 1). When an image is labeled negative, we assume no car is present in the scene and, hence, the labels of the features are set to the negative class. Otherwise, when an image contains at least one car, we do not know whether individual points correspond to car or background, so learning implicates determining the feature labels. We call this approach *semi-supervised learning by data association*. Each label is a binary variable that determines whether or not a feature belongs to the object. Since the training data does not contain any positively labeled regions, we introduce label constraints to learn which features belong to cars. The constraints introduce a new set of parameters, but we show that our model's performance is reasonably insensitive to these parameters, and in fact performs quite well with weak constraints. Note we can achieve multi-category classification by combining responses from multiple binary classifiers [31].

One might be skeptical that one can successfully learn to recognise objects in scenes from such weakly labeled data, given the high dimension of the features, the wide variability exhibited in the training scenes, and the fact that there are often as many as a thousand unlabeled points per image! Recent publications demonstrate good performance in similar but less ambitious tasks by using statistical translation models [5, 6, 10], support vector machines [2, 3] and sparse Bayesian kernel machines [15]. Other publications use captioned images in order to learn representations of objects [1, 9, 11], but none of them explicitly perform data association. Instead, they treat unlabeled background features as noise (in other words, an inconvenience). While difficult to verify independently, we argue that variable selection is more effective when explicitly modeling the classification of data points, because it allows the model to exploit unlabeled background features.

We employ an augmented Bayesian classification model with an efficient Markov Chain Monte Carlo (MCMC) algorithm [15] to simultaneously learn the unobserved labels and select a sparse object class representation from the extracted high-dimensional descriptors. We introduce a generalised Gibbs sampler to explore the space of labels that satisfy the constraints. Bayesian learning comprehends approximation of

the posterior distribution through integration of multiple hypotheses, not only a crucial ingredient for robust performance in noisy environments, but also resolves sensitivity to initialisation. Expectation Maximisation or other data augmentation gradient descent methods approximate the distribution with a single mode, leading to a misrepresentation the true model posterior. For practical MCMC exploration of the posterior's modes, however, the posterior must be sufficiently peaked. This consideration exposes a major problem with mixture models — motivating a Bayesian kernel model — since the number of modes explodes factorially relative to the number of components [7]. An additional advantage over cited methods is that we do not need to reduce the dimension of the descriptors through unsupervised techniques, such as vector quantization or principal component analysis, which may purge valuable information.

The speed at which our blocked Gibbs sampler explores the space of likely solutions is critical to success of the algorithm. We improve our MCMC algorithm by introducing a latent parameter which decouples highly correlated variables and thus leads to improved convergence, as suggested in [19]. Experiments on large data sets indicate that the proposed parameter expanded and blocked Gibbs sampler converges to posterior regions of high probability after a few thousand iterations. The main computational curse harboured by our MCMC algorithm is one shared with all kernel machine methods: learning complexity is $O(N^2)$, where N is the number of data points in the training set. Since our experiment data often includes hundreds of thousands of points in high dimensions, a theme in this article and in future work is reducing the expense of Bayesian learning.

The following section outlines the Bayesian model for semi-supervised classification. Sec. 3 details how we design a blocked Gibbs sampler to resolve model learning. In Sec. 4, we describe our implementation of existing vision techniques for the extraction and representation of candidate features. Finally, we present experimental results on object recognition tasks in Sec. 5 and offer conclusions in Sec. 6.

2 Model Specification

The training data consists of the set of labeled images, or documents, $d = \{d_1, d_2, \ldots, d_D\}$. We represent image d_j by a set of feature vectors $\{x_i \,|\, i \in d_j\}$, denoting the entire set of N training features by $x \triangleq \{x_1, x_2, \ldots, x_N\}$. We connote an observed label by y_i^k and an unknown label by y_i^u, where $y_i \in \{0, 1\}$. When the image is labeled as being negative, the labels of the points are observed to be $y_i^k = 0$. When a document label is positive, we sample the labels y_i^u keeping in mind some additional constraints on the number of labels per class. Label constraints can play a crucial role in learning the unknown labels, most notably when we do not have any known labels of a certain class, such as when all the training images positive images also possess feature points from other objects (e.g. see Fig. 1). We define n_0 to be the constraint on the minimum number of negative points in an unlabeled document, and n_1 to be the minimum number of positively classified points.

We adopt a probit link classifier

$$y_i = \begin{cases} 1 & \text{if } f(x_i, \beta, \gamma) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

with $f(x_i, \beta, \gamma)$ specified below. By convention, researchers tend to adopt a logistic (sigmoidal) link function, but from a Bayesian computational point of view, the probit link has many advantages and is equally valid. Following Tham *et al.* [30], the unknown function is a sparse kernel machine:

$$f(x_i, \beta, \gamma) = \beta_0 + \sum_{k=1}^{K} \gamma_k \beta_k \psi(x_i, x_k), \tag{2.2}$$

where ψ denotes the kernel function and K is the number of kernels. Usually we assign the kernels to be the set of data points, but we keep the notation separate to retain generality of the model. To avoid

confusion, we subscript all data points with i and all kernel centres with k. We use the Gaussian kernel $\psi(x_i, x_k) = \exp(-\lambda(x_i - x_k)^2)$, but other choices are possible. As with all kernel methods, successful classification depends on a good choice on the scale, or bandwidth, λ.

The sparsity of support vectors machines comes directly as a result of the objective function. Such is not the case in the Bayesian framework. We artificially introduce sparsity through a set of binary variables, $\gamma \triangleq [\gamma_1, \gamma_2, \ldots, \gamma_N]$, that are set to 1 when the respective kernel centres are active, and otherwise they are 0. We denote the vector of regression coefficients by $\beta \triangleq [\beta_0, \beta_1, \ldots, \beta_N]^T$.

By adopting a set of one-dimensional independent variables $z \triangleq \{z_1, z_2, \ldots, z_N\}$ with distribution

$$p(z_i \,|\, \gamma, \beta, x_i) = \mathcal{N}(\Psi_{\gamma i}\beta, 1), \tag{2.3}$$

we can analytically compute the posterior of the high-dimensional coefficients β within the standard linear-Gaussian model, a strategy first introduced by Nobel Laureate Daniel McFadden [23]. Ψ_γ is the $N \times K$ kernel response matrix with zeroed columns corresponding to inactive entries of γ. Ψ_i denotes the ith row of the kernel response matrix, and is given by

$$\Psi_i \triangleq \left[\begin{array}{cccc} \psi(x_i, x_1) & \psi(x_i, x_2) & \ldots & \psi(x_i, x_K) \end{array} \right].$$

We follow a hierarchical Bayesian strategy, where the unknown parameters and labels are drawn from appropriate prior distributions. The intuition behind this hierarchical approach is that by increasing the levels of inference, we can make the higher level priors increasingly more diffuse. That is, we avoid having to specify sensitive parameters and, therefore, are more likely to obtain results that are independent of parameter tuning. We assume the regression coefficients are normally distributed with zero mean, and covariance S scaled by learned parameter δ^2:

$$p(\beta \,|\, \gamma, \delta^2) \;=\; \mathcal{N}(\beta \,|\, 0, \delta^2 S_\gamma). \tag{2.4}$$

At this point, we leave the choice of S open and later we derive results using specific choices for the prior covariance term. In the future, we would like to learn a parameterised covariance over the regression coefficients, such as the Relevance Vector Machine [31] which learns a weight for every β_k. S_γ refers to the covariance matrix whereby the elements of a row and column corresponding to an inactive kernel k is set to 0, and correspondingly the β_k is shrunk to zero. We assign an inverse Gamma to the scale parameter δ^2,

$$p(\delta^2 \,|\, \mu, \nu) \;=\; \mathcal{IG}(\delta^2 \,|\, \tfrac{\mu}{2}, \tfrac{\nu}{2}). \tag{2.5}$$

μ, ν are fixed hyperparameters typically set to near-uninformative values.

Each γ_k follows a Bernoulli distribution with success rate $\tau \in [0, 1]$:

$$p(\gamma \,|\, \tau) \;=\; \tau^{\Sigma\gamma}(1 - \tau)^{K - \Sigma\gamma}, \tag{2.6}$$

where we define $\Sigma\gamma \triangleq \sum_{k=1}^{K} \gamma_k$ to be the number of active kernels. Rather than fix the success rate, we have τ follow a Beta distribution with parameters $a, b \geq 1$. This lets the model to adapt to the data while allowing the user some control over the prior. We integrate out the nuisance parameter τ to obtain the prior

$$\begin{aligned} p(\gamma \,|\, a, b) \;&=\; \int_0^1 p(\gamma \,|\, \tau)\, p(\tau \,|\, a, b)\, d\tau \\ &\propto \int_0^1 \tau^{\Sigma\gamma + a - 1} \left(1 - \tau\right)^{K - \Sigma\gamma + b - 1} d\tau \\ &=\; \frac{\Gamma(\Sigma\gamma + a)\, \Gamma(K - \Sigma\gamma + b)}{\Gamma(K + a + b)}\,. \end{aligned} \tag{2.7}$$

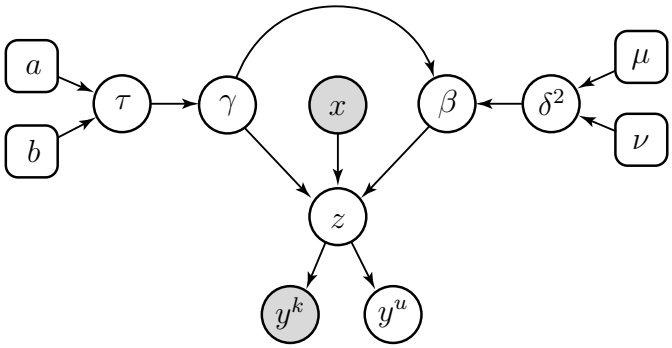


Figure 2: The directed graphical representation of the semi-supervised classification model. Shaded nodes are observed and square nodes are fixed hyperparameters. We do not show the dependencies between the y_i^u's.

Small values of the mean $a/(a+b)$ skews our prior belief towards a sparse model. The term $a+b$ controls our confidence in such a prior. By setting $b \gg a$ on large data sets, we can initialise the Gibbs sampler to a feasible (small) number of active kernels.

The regression coefficients z_i^u in document d_j are not independent of each other, so we need to rectify (2.3) with the joint expression

$$p(\{z_i^u\} \,|\, \gamma, \beta, \{x_i\}) \;\propto\; \prod_i \mathcal{N}(z_i^u \,|\, \Psi_{\gamma i}\beta, 1)\, \mathbb{I}_{C_0}(\{z_i^u\})\, \mathbb{I}_{C_1}(\{z_i^u\}) \qquad \text{such that } i \in d_j, \tag{2.8}$$

and where C_0 is the set of assignments to y_i^u (and accordingly z_i^u) that obey the negative labels constraint n_0, C_1 is the set of assignments to y_i^u that satisfy the constraint on the minimum number of positive labels n_1, and $\mathbb{I}_\Omega(\omega)$ is the set indicator function: 1 if $\omega \in \Omega$, and 0 otherwise.

In summary, the fixed parameters of the model are $\{\lambda, a, b, \mu, \nu, n_0, n_1\}$, the model parameters we retain for predictions are $\theta \triangleq \{\gamma, \beta\}$ and the learned nuisance parameters are $\{\delta^2, \tau\}$. The directed graphical representation of the hierarchical model is shown in Fig. 2.

3 Model Computation

The classification objective is to estimate the label y_{N+1} of a new unseen point x_{N+1} given the training data,

$$p(y_{N+1} = 1 \,|\, x_{N+1}, x, y^k) \;=\; \int p(y_{N+1} = 1 \,|\, x_{N+1}, \theta)\, p(\theta \,|\, x, y^k)\, d\theta. \tag{3.1a}$$

The integral is intractable, so we approximate (3.1a) by

$$\begin{aligned} p(y_{N+1} = 1 \,|\, x_{N+1}, x, y^k) \;&\approx\; \frac{1}{n_s} \sum_{s=1}^{n_s} p\big(y_{N+1} = 1 \,|\, x_{N+1}, \theta^{(s)}\big) \\ &=\; 1 - \frac{1}{n_s} \sum_{s=1}^{n_s} \Phi\big(-\Psi_{\gamma^{(s)}_{N+1}} \beta^{(s)}\big), \end{aligned} \tag{3.1b}$$

where n_s is the number of samples generated, θ is the set of parameters that directly influence prediction, $\Phi(\cdot)$ is the univariate Normal cumulative density function, and each sample $\theta^{(s)} = \{\gamma^{(s)}, \beta^{(s)}\}$ follows the posterior $p(\theta \,|\, x, y^k)$. We spend the rest of this section designing an MCMC algorithm that approximates the posterior distribution. Since the algorithm is quite involved, we break up the derivation into several sections.

In Sec. 3.1, we derive the blocked Gibbs sampler, with the exception of the sampler for γ, left to Sec. 3.3 since it is an elaborate step. Sec. 3.2 improves upon the expected convergence rate of the Gibbs sampler through introduction of an expansion parameter.

3.1 The Blocked Gibbs sampler

Kück *et al.* [15] develop an MCMC algorithm for sampling from the posterior by augmenting the original blocked Gibbs sampler [30] to the data association scenario. As in [15], we sample the high-dimensional parameters β and the regularisation parameters directly from their posterior distributions.

If two variables are highly correlated, an ordinary Gibbs sampler will slowly navigate through the joint space because it samples from the full conditional distributions. We will see shortly that we can analytically derive the joint of γ and β, resulting in so-called "blocked" moves which converge faster than a Markov chain derived from the full conditionals [18]. We factorise the joint conditional of the regression coefficients and variable selection parameters as

$$p(\beta, \gamma \,|\, \delta^2, a, b, x, z) \;=\; p(\beta \,|\, \gamma, \delta^2, z, x)\, p(\gamma \,|\, a, b, \delta^2, z, x). \tag{3.2}$$

We obtain β from the conditional posterior,

$$p(\beta \,|\, \gamma, \delta^2, z, x) \;=\; \mathcal{N}\left(\beta \;\middle|\; Q_\gamma \Psi_\gamma^T Z, Q_\gamma\right). \tag{3.3}$$

where $Q \triangleq (\Psi^T\Psi + (\delta^2 S)^{-1})^{-1}$. See Sec. C.1 for the derivation. Beware that the order of matrix multiplication implied in (3.3) is not efficient. In fact, the time complexity for sampling β is $O(K^2)$ by taking advantage of values previously computed during sampling of the γ parameter (see Sec. 3.3). We leave the details for sampling γ to Sec. 3.3.

We can sample the z_i^k's easily since they are independent of each other. Using Bayes' theorem, the posterior distribution for each z_i^k is given by

$$\begin{aligned} p(z_i^k \,|\, y_i^k, \gamma, \beta, x_i) \;&\propto\; p(y_i^k \,|\, z_i^k)\, p(z_i^k \,|\, \gamma, \beta, x_i) \\ &= \begin{cases} \mathcal{N}(\Psi_{\gamma i}\beta, 1)\, \mathbb{I}_{(0,+\infty)}(z_i^k) & \text{if } y_i^k = 1 \\ \mathcal{N}(\Psi_{\gamma i}\beta, 1)\, \mathbb{I}_{(-\infty,0]}(z_i^k) & \text{otherwise.} \end{cases} \end{aligned} \tag{3.4}$$

Sampling the z_i^u's is not quite as simple because the joint posterior does not factorise. While [15] use rejection sampling to sample the unknown labels subject to the constraints, we adopt a more efficient MCMC scheme and sample from the full conditionals in each document:

$$p(z_i^u \,|\, \{z_{\neq i}^u\}, \gamma, \beta, x_i) \;=\; \begin{cases} \mathcal{N}(\Psi_{\gamma i}\beta, 1)\, \mathbb{I}_{(0,+\infty)}(z_i^u) & \text{if } \mathbb{I}_{C_1}(\{z_{\neq i}^u\}) = 0 \\ \mathcal{N}(\Psi_{\gamma i}\beta, 1)\, \mathbb{I}_{(-\infty,0]}(z_i^u) & \text{if } \mathbb{I}_{C_0}(\{z_{\neq i}^u\}) = 0 \\ \mathcal{N}(\Psi_{\gamma i}\beta, 1) & \text{otherwise,} \end{cases} \tag{3.5}$$

such that $i \in d$, and $\neq i$ is defined to be the set $\{i' \,|\, i' \neq i \text{ and } i' \in d\}$. Note that C_0 and C_1 are never both unsatisfied at the same time.

We sample the regression coefficients scale δ^2 from the conditional posterior

$$\begin{aligned} p(\delta^2 \,|\, \gamma, \beta, \mu, \nu) \;&\propto\; p(\beta \,|\, \gamma, \delta^2)\, p(\delta^2 \,|\, \mu, \nu) \\ &= \mathcal{IG}\left(\tfrac{1}{2}(\mu + \Sigma\gamma), \tfrac{1}{2}\left(\nu + \beta^T S_\gamma^{-1}\beta\right)\right). \end{aligned} \tag{3.6}$$

3.2 Parameter expansion using an inverse Gamma prior

The convergence rate of the Gibbs sampler suffers from high correlation between parameter β and the latent variables z [18]. We introduce an auxiliary variable α which overparameterizes the model by scaling z, but also increases the variance of β given z, leading to larger transitions in the Markov chain. We define the transformation on z_i by $t_\alpha = z_i/\alpha$, and the according Jacobian is $J_\alpha(z) = \det\{\partial t_\alpha(z_i)/\partial z_i\} = \alpha^{-N}$. Liu *et al.* [19] suggest placing an improper Haar prior on α since it is optimal for convergence. In practice, however, the Haar prior tends to be unstable. An alternative is the inverse Gamma prior, such that $\alpha^2 \sim \mathcal{IG}(\frac{\mu_\alpha}{2}, \frac{\nu_\alpha}{2})$. It achieves improved an convergence rate while allowing the user to tune μ_α and ν_α for stability.

From the results derived in Sec. C.2, the parameter expanded Gibbs sampler for β and z consists of the following steps:

1	Draw z	$\sim p(\cdot \mid y, \gamma, \beta, x_i)$	$\triangleright$ See equations 3.4 and 3.5.
2	Draw α_0^2	$\sim p(\cdot \mid \mu_\alpha, \nu_\alpha)$	$\triangleright$ See equation 3.7.
3	Draw α^2	$\sim p(\cdot \mid \alpha_0 z, \gamma, \beta, \mu_\alpha, \nu_\alpha)$	$\triangleright$ See equation C.4.
4	z'	$\leftarrow \sqrt{\alpha_0/\alpha} \times z$	
5	Draw β	$\sim p(\cdot \mid \gamma, \delta^2, z', x)$	$\triangleright$ See equation 3.3.

The sampling distribution for α^2 in Step 3 can be computed in time $O(K^3)$ by using previously computed values, as for β. We omit the other Gibbs sampler steps since they are not affected by the inclusion of the expansion parameter. To sample α_0 from the prior distribution $p(\alpha \mid \mu_\alpha, \nu_\alpha)$, we note that

$$p(\alpha \mid \mu_\alpha, \nu_\alpha) \propto \mathcal{IG}\big(\tfrac{\mu_\alpha+1}{2}, \tfrac{\nu_\alpha}{2}\big) \times \tfrac{d\alpha^2}{d\alpha}. \tag{3.7}$$

3.3 Sampling the variable selection parameter

Exact sampling from the γ posterior is impractical because it requires computation of the 2^K possible assignments of γ. We elaborate on an alternative first proposed in [30]. It is essentially a Gibbs sampler implemented using a Metropolis-Hastings (M-H) proposal for more efficient computation. While the Metropolised Gibbs sampler achieves satisfactory acceptance rates, its complexity is still $O(NK^2)$, since it has to cycle through all the K kernels in the worst case. It is worth our while to tune computation to the choice of S, so we assume a stabilised g-prior $S = (\Psi^T\Psi + \epsilon I_K)^{-1}$, where I_K is the $K \times K$ identity matrix and ϵ is a positive number generally chosen to be much smaller than the values contained in the Gram matrix $\Psi^T\Psi$. The stabilisation term helps maintain a prior covariance with full rank.

We observe that sampling from the posterior of γ is impractical, but it is possible to sample each γ_k from its full conditional — that is, the distribution with all the other kernel centres fixed. The key observation made by Tham [30] is that we can improve computation by formulating each Gibbs draw as a Metropolis-Hastings step [17]. By manipulating some of the matrix computation, we present here an algorithm with complexity $O(NK^2)$ with $K \leq N$, instead of $O(K^2(N + K^2))$ assumed by naively multiplying and inverting matrices. On our data sets, this gives us significant computational savings.

The main idea behind the Metropolis-Hastings algorithm is to sample a candidate value from a proposal distribution $q(\gamma^\star \mid \gamma^{(i)})$, where we denote the ith sample with the superscript (i), and then accept the candidate $\gamma^{(i+1)} = \gamma^\star$ with probability $\mathcal{A}(\gamma^{(i)}, \gamma^\star)$. Otherwise, the new sample remains unchanged and $\gamma^{(i+1)} = \gamma^{(i)}$. The acceptance probability is given by

$$\mathcal{A}(\gamma^{(i)}, \gamma^\star) = \min\left\{1, \frac{p(\gamma^\star \mid \delta^2, a, b, x, z)\, q(\gamma^{(i)} \mid \gamma^\star)}{p(\gamma \mid \delta^2, a, b, x, z)\, q(\gamma^\star \mid \gamma^{(i)})}\right\}. \tag{3.8}$$

It is generally difficult to design a sampling distribution $q(\gamma^\star \,|\, \gamma^{(i)})$ that permits a high acceptance rate. In our case, $q(\gamma^\star \,|\, \gamma^{(i)})$ derives directly from the full conditional of γ_k and consequently our acceptance rate is guaranteed to be high. We direct the reader to a couple excellent introductions to the M-H algorithm [4, 8].

We need to consider two cases for the Metropolised Gibbs sampler: when $\gamma_k = 0$ and when $\gamma_k = 1$. When kernel k is inactive, our proposal consists of flipping γ_k to 1 with probability $p(\gamma_k^\star = 1 \,|\, \gamma_{\neq k}, a, b) \propto p(\gamma_k^\star = 1, \gamma_{\neq k}, a, b)$ (see equation C.3a). When $\gamma_k = 1$, the acceptance probability is

$$\begin{aligned}\mathcal{A}\big(\gamma_k^\star = 1, \gamma_k^{(i)} = 0\big) &= \min\left\{1, \frac{p(z \,|\, \gamma_k^\star = 1, \gamma_{\neq k}, \delta^2, x)}{p(z \,|\, \gamma_k^{(i)} = 0, \gamma_{\neq k}, \delta^2, x)}\right\} \\ &= \min\left\{1, \sqrt{\xi_k^\star \left(\delta^2 \zeta_k^\star\right)^{-1} \exp\left(\xi_k^\star \left(Z^T \left(I_N - v\Psi_{\gamma^{(i)}} Q_{\gamma^{(i)}} \Psi_{\gamma^{(i)}}^T\right) \Psi_{\gamma_k^\star}\right)^2\right)}\right\} \qquad (3.9a)\end{aligned}$$

following from the derivation of (C.6) in Sec. C.3. We define $\Psi_{\gamma_k^\star}$ to be the column vector of the kernel response matrix introduced by the new kernel k, $v \triangleq \frac{1+\delta^2}{\delta^2}$, $\zeta_k \triangleq (\epsilon + v\Psi_{\gamma_k}^T (I_N - \Psi_0 S_0 \Psi_0^T)\Psi_{\gamma_k})^{-1}$ and $\xi_k \triangleq (\frac{\epsilon}{\delta^2} + v\Psi_{\gamma_k}^T (I_N - v\Psi Q \Psi^T)\Psi_{\gamma_k})^{-1}$. There is always a subscript $\gamma^{(i)}$ on the variables Ψ, S and Q. When kernel k is active, we deactivate it with probability $p(\gamma_k^\star = 0 \,|\, \gamma_{\neq k}, a, b)$ (see equation (C.3b)), and then accept the change with probability

$$\begin{aligned}\mathcal{A}\big(\gamma_k^\star = 0, \gamma_k^{(i)} = 1\big) &= \min\left\{1, \frac{p(z \,|\, \gamma_k^\star = 0, \gamma_{\neq k}, \delta^2, x)}{p(z \,|\, \gamma_k^{(i)} = 1, \gamma_{\neq k}, \delta^2, x)}\right\} \\ &= \min\left\{1, \sqrt{\delta^2 \zeta_k^{(i)} \left(\xi_k^{(i)}\right)^{-1} \exp\left(-\xi_k^{(i)} \left(Z^T \left(I_N - v\Psi_{\gamma^\star} Q_{\gamma^\star} \Psi_{\gamma^\star}^T\right) \Psi_{\gamma_k^{(i)}}\right)^2\right)}\right\}. \qquad (3.9b)\end{aligned}$$

Depending on the strength of the γ prior, the Metropolised Gibbs sampler can filter out a lot of poor candidates while maintaining a desirable M-H acceptance rate. See Appendix D for an efficient implementation of the sampler.

4 Selection of local features

The first step in our object recognition approach is to employ a detector (or several) to filter out regions of interest and then compute a rotation-invariant descriptor for each region. We use four different detectors for generating observations x in a scene. The Harris-Laplace detector [24] finds corner-like features, the Kadir-Brady detector [14] proposes circular regions with histograms of grey-level maximum entropy, and the Laplacian of Gaussian (LoG) [16] and Difference of Gaussians (DoG) [21] — the latter an approximation to the former — focus on regions of uniform intensity. The extracted regions are mapped to discs with a fixed radius in order to achieve scale invariance. In some experiments, we use an affine-invariant version of the Harris-Laplace detector. We use the procedure described in [25] to produce features invariant to rotation and affine illumination changes. Based on earlier studies of object matching at different viewpoints and illumination changes [26], we choose SIFT [21] to describe the normalised regions extracted by the detectors. We compute each SIFT descriptor using 8 orientations and a 4×4 grid, resulting in a 128-dimension feature vector.

5 Experiments

The goal is to design experiments that assess our model's capacity for recognising objects in unseen images. We evaluate recognition through image classification — identifying the presence or absence of objects in

Figure 3: Images from the arctic dogs data set. The two leftmost images are labeled positive because they contain instances of dogs. The two on the right are background scenes.

images — because it is a well-defined problem. We rank images for classification by summing the feature label probabilities $p(y_i = 1 \mid x_i)$ assigned by the model. Unless the image data is well-constructed, however, it is hard to argue that image classification equates object recognition. We want to make sure we are learning to recognise cars, not objects associated with cars, such as stop signs. We address these concerns by creating three new data sets, where all the images emanate from the same environment: parking lots with and without cars, and arctic landscapes with and without dogs and polar bears. All three sets exhibit a significant amount of variation in scale, pose and lighting conditions. Examples of cars and dogs can be found in Figures 1 and 3, respectively. For purposes of comparison with other methods, we also show results on existing data sets. We summarize the data used in our experiments in Table 1.

	Training images		Test images	
Object class	with object	background	with object	background
airplanes	400	450	400	450
motorbikes	400	450	400	450
wildcats	100	450	100	450
bikes	100	100	50	50
people	100	100	50	50
cars	50	50	29	21
arctic dogs	26	36	13	19
polar bears	19	43	9	23

Table 1: Summary of the the experiment data. The airplanes, motorbikes and wildcats originate from [11], and the bikes and people were first used in [29]. The dogs and polar bears come from the Corel data set. The car data set is available by request.

	H-L	K-B	LoG	DoG	Affine H-L	Combo	Random	Fergus *et al.*	Opelt *et al.*
airplanes	0.985	0.993	0.938	—	—	0.998	—	0.902	0.889
motorbikes	0.988	0.998	0.983	—	—	1.000	—	0.925	0.922
wildcats	0.960	0.980	0.930	—	—	0.990	—	0.900	—
bikes	0.920	0.880	0.840	0.860	0.880	0.900	0.920	—	0.865
people	0.800	0.740	0.840	0.840	0.700	0.820	0.800	—	0.808
cars	0.966	0.897	0.897	0.897	0.931	0.931	0.690	—	—
arctic dogs	0.615	0.538	0.769	0.615	0.692	0.615	0.462	—	—
polar bears	0.667	0.556	0.556	0.556	0.556	0.556	0.667	—	—

Table 2: Image classification performance on test sets measured using the Receiver Operating Characteristic (ROC) equal error rate. The last two columns refer to the performance reported by Fergus *et al.* [11] and Opelt *et al.* [29]. All the other columns state the performance obtained using our proposed Bayesian model with regions extracted from various detectors (from left to right): Harris-Laplace [24], entropy detector proposed by Kadir and Brady [14], Laplacian of Gaussians [16], Difference of Gaussians [21], affine Harris-Laplace [25], combination of the Harris-Laplace, Kadir-Brady and LoG, and regions randomly selected from all possible positions and scales in each image.

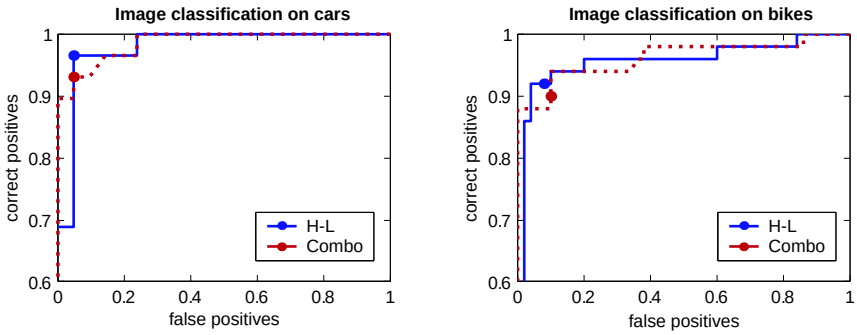


Figure 4: The graph on the left plots the ROC curve for classification performance of car test images using the Harris-Laplace detector (blue solid line) and the combination of three detectors (red dotted line). The graph on the right shows analogous results for the bike test set. In both cases, the equal error rate (indicated by a large dot) is inferior in the combination, but upon closer inspection, one could justifiably argue (depending on the error measure used) that the combination performs slightly better.

Table 2 reports image classification results on all data sets. Before we enter into a discussion of our results, we first explain how we conduct the experiments. For fair comparison, we adjust the thresholds of the detectors in order to obtain an average of 100 detections per training image. Therefore, the combination scenario (the sixth column in Table 2) possesses an average of 300 detections per image. We also include results using a random detector, which indiscriminately selects a subset of 100 features from the entire collection of regions at all scales and locations. We compare performance on the airplanes, motorbikes, wildcats, bikes and people with the results obtained in [11, 29], but we do not control for the numbers of detections: Fergus *et al.* [11] extract only 20 features per image on average, in part owing to the poor time complexity of their method, while Opelt *et al.* [29] have a distinct advantage since they learn from several hundred regions per image.

In all our experiments, we fix the label constraint n_0 to 0 and set n_1 between 15 and 30, depending on the object in question. Our constraints tend to be conservative, the advantage being that they do not force too many points to belong to objects that occupy only a small portion of the scene. (In a series of separate trials on the cars, we varied n_1 between 10 and 100 and observed no identifiable relationship between image classification performance and constraint strength.) We set $a = 1$ and b conforming to a variable selection prior of approximately 200 active kernel centres. We bestow near uninformative priors on the rest of the model parameters. We find that 2000 MCMC samples with a burn-in period of 100 is sufficient for a good approximation of the model posterior. As previously noted in this paper and confirmed in independent trials, the Gaussian kernel scale parameter λ has a significant impact on the success of the learning algorithm. Large values of λ — scales that concentrate the kernel mass to small neighbourhoods — do not generalise well. High or diffuse scales appear to work best, which makes sense because they introduce uncertainty into the kernel space, a boon for noisy tasks. However, large scales render our learning algorithm unstable, and in experiments we notice that the Markov chain often fails to converge to a common posterior distribution. In all our experiments, we set λ to $1/100$ because our MCMC algorithm reliably converges to a good solution. We note that scale selection is an unsolved problem.

We measure performance with the Receiver Operating Characteristic (ROC) equal error rate, since it is a standard evaluation criterion [11, 29]. It is defined to be the point on the ROC curve — obtained by varying the classification threshold — when the proportion of true positives is equal to the proportion of true negatives.

We now highlight and discuss some of the more interesting results of Table 2. First and most significantly, we observe that our model in combination with the three detectors always provides a better image classification than [11, 29]. Independent MCMC trials exhibit little variance, so we can state confidently

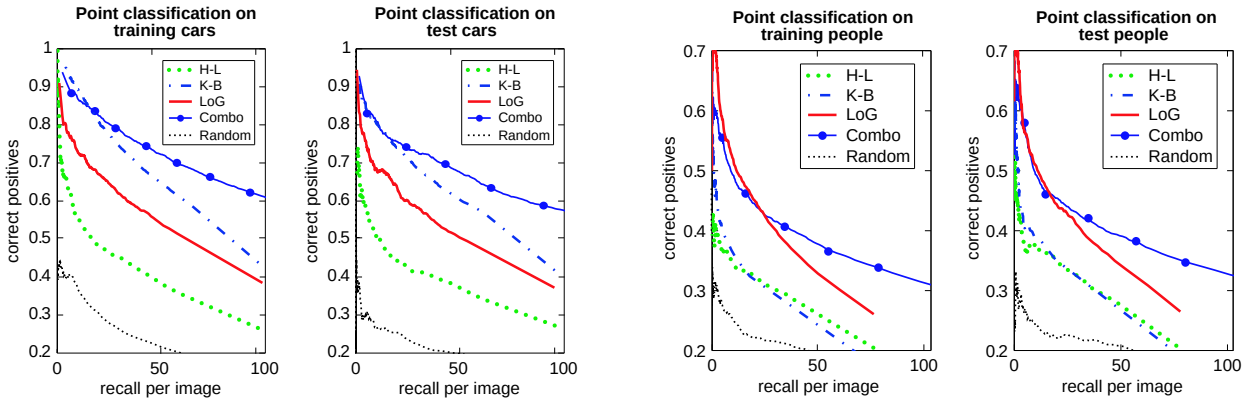


Figure 5: Plots of precision (percentage of correct positives) versus average recall per image for the task of labeling individual features on cars and people. For reference, we show performance of the random detector, indicated by the dotted black line. Note that the Combo curve continues to the 300 mark in recall per image. Our algorithm learns which features are best in the combination, but this performance does not necessarily translate to better image classification in Table 2.

that our improvements are well outside standard deviation intervals. Intriguingly, no detector dominates classification of all the eight object classes. The combination of the Harris-Laplace, Kadir-Brady and LoG detectors often, albeit inconsistently, improves the equal error rate. Further analysis of the performance on the cars and bikes using the Harris-Laplace detector and the combination in Fig. 4 shows that the equal error rate measure can be deceptive: the ROC curves imply that the combination of detectors is an improvement, a result otherwise indicated by the equal error rate. From Fig. 5 (plot of precision versus average recall per image for classification of individual features), we observe that the combined classifier learns which interest regions are best, even though the combination does not necessarily translate to better image classification performance (e.g. for people). Fig. 6 shows a couple examples where the combination results in a better classification than the individual detectors.

The discrepancy between image classification performance in Table 2 and feature classification performance in Fig. 5 shows that there exists an experimental gap between the two tasks. For instance, Harris-Laplace features are poorly identified as cars (see Fig. 5), but they work best for determining whether a car is present or absent in a scene (see Table 2). The implication is that most Harris-Laplace features are indiscriminative on their own, but a small subset of the extracted regions are indeed very useful for image classification. Likewise, random features compose a reasonably good classifier for people images even though they are not discriminative individually, as indicated by the plots in Fig. 5. The difference between random features and Harris-Laplace features, however, is that we can account for poor classification of individual random features because the model learns trends in the background of positive images. This goes on to explain why random features are unsatisfactory for object detection in the car data set, since the background is the similar in positive and negative images. Examples of correctly classified images in Fig. 7 show that the model learns to associate background regions with people.

Still, random features often do very well for image classification. What can we make of this surprising result? It shows that it is often difficult to extract *a priori* useful object features. This is particularly the case for people, since exhibit great inter-class variability. Furthermore, the descriptors we employ are not appropriate for certain object classes, including bikes (shape features, such as those proposed in [13], might work better). Cars, on the other hand, exhibit less variability within the same class, and consequently our detectors consistently propose good features (wheels, corners of windows), which explains in part why random features perform comparitively poorly. The other reason is that the cars (and dogs and polar bears)

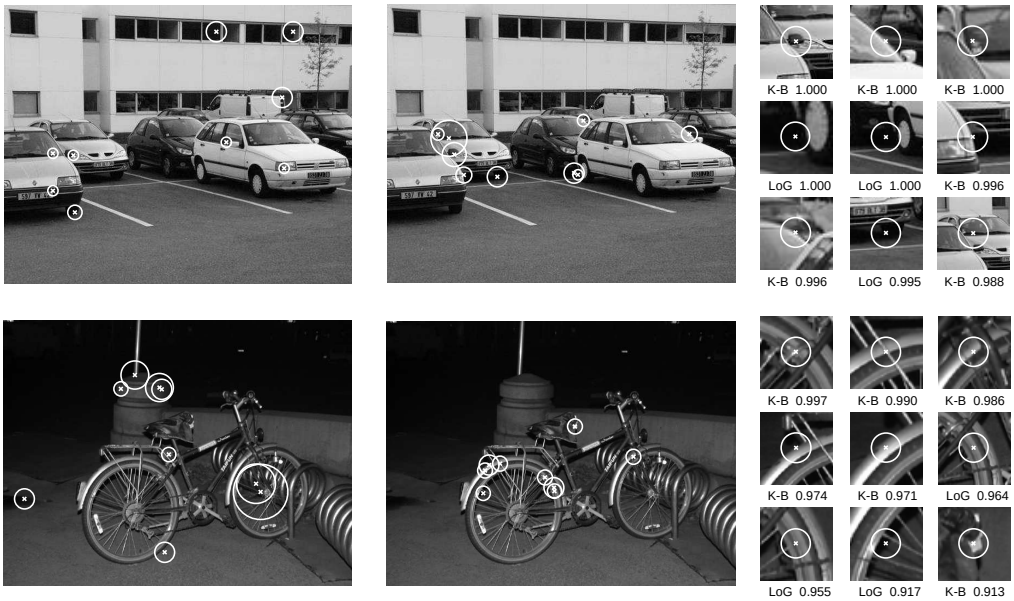


Figure 6: Two examples in which the combination results in an improved classification. The circles represent the 9 Harris-Laplace (left column) and combination (middle column) interest regions that are most likely to belong to cars or bikes. On the right, we display the top features along with feature type and probability of positive classification.

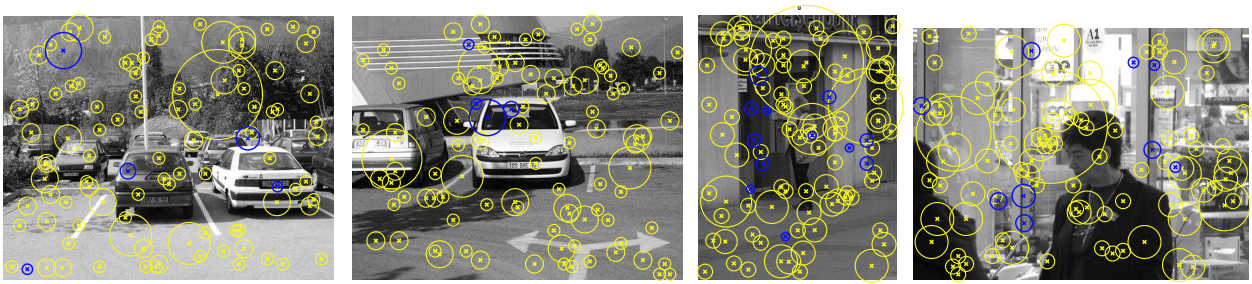


Figure 7: Examples of correctly classified images using randomly extracted interest regions. The circles depict all the random regions along with their characteristic scale. Dark blue circles represent regions that are more likely than not to belong to the object, while light yellow circles more probably belong to the background. The captions indicate the correct labels. The model learns to associate background regions with people.

often occupy a small portion of the scene, necessitating a reasonable level of detector precision.

Another interesting result: in most cases, the affine-invariant transformation of the Harris-Laplace detector reduces recognition ability, hinting that valuable information is lost in the description.

Finally, we wondered how increasing the level of supervision by adding labels to extracted regions would affect performance. The results of our experiments on cars using the Harris-Laplace detector and arctic dogs using the Laplacian of Gaussian detector are shown in Fig. 8. The ROC curves quantify the models' accuracy in labeling individual image features. As expected, the addition of a few hand-labeled points improves recognition on the car training set. However, further upgrades in supervision result in almost no gains in generalisation to the car test images. These results suggest that there is nearly enough supervised information in the car captions to allow the model to correctly label and select the most reliable features in unseen images. On the other hand, our model profits much more from supervision for recognising dogs, presumably because there is not nearly enough information to properly learn the data association. We show some examples of recognising cars and dogs with different levels of training supervision in Fig. 9.

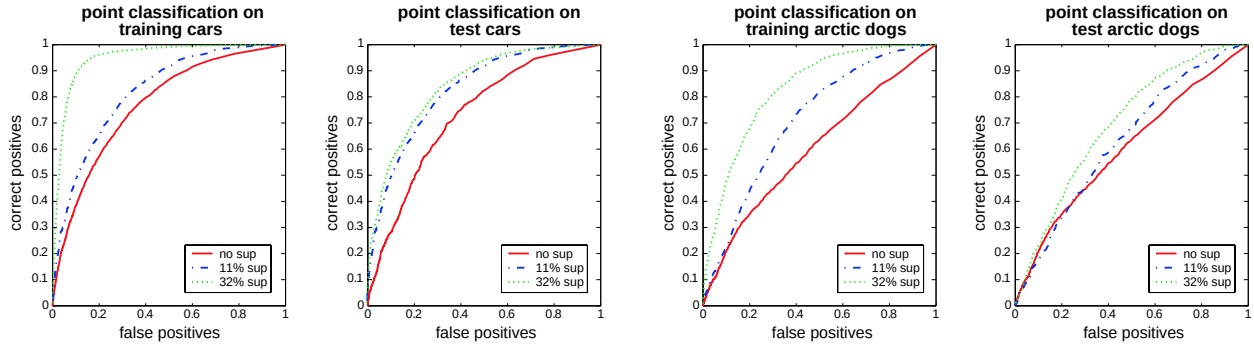
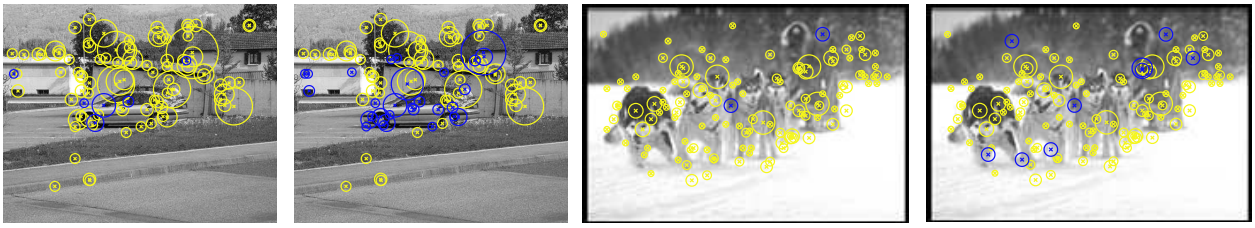


Figure 8: The ROC plots demonstrate how learning with different proportions of hand-labeled points affects performance of labeling individual features. We use the Harris-Laplace detector for the cars, and the LoG for the dogs.



(a) contains cars (b) contains cars (c) contains dogs (d) contains dogs

Figure 9: Classification of individual interest regions using classifiers trained with various levels of supervision (see Fig. 8). We show all the extracted interest regions along with their characteristic scale: dark blue circles are more likely than not to belong to the object, while light yellow circles more probably belong to the background. *(a)* Car test image, no observed car labels during training. *(b)* The same image, except that the model trained with 11% labels in positive images. *(c)* and *(d)* show dog label predictions, without and with additional supervision (11%), respectively. The caption below each image is the correct label.

To complete this section, we show some examples of correctly and incorrectly classified images in Figures 10 and 11. The figures show all the regions extracted by the detectors.

6 Conclusions

In this paper, we extended the discriminative power of local scale-invariant features using Bayesian learning. Our method allows us to solve the important problem of selecting and combining multiple local features. We showed that constrained semi-supervised learning using MCMC is remarkably well-behaved in the face of noisy high-dimensional features and wide variability in the unlabeled training data. On a less encouraging note, our results indicate that there remains a lot of work ahead on proposing initial detections as a basis for object representations, because random SIFT features are often competitive with current detection schemes.

The model we propose is very general, and it can be easily extended to compare the discriminative ability of a variety of features, including shape and texture. One future extension to our learning framework is the incorporation of soft constraints through Bayesian priors, which adapt the number of positive labels according to the size of the object relative to the the background. An obvious but nonetheless interesting extension is multi-category classification. We believe that are there still some intriguing opportunities for improving the expected complexity of computation using ball trees [20]. Finally, we note that while we propose new data sets that accurately evaluate object recognition, an neglected but important direction of research is learning of vision models through user reinforcement.

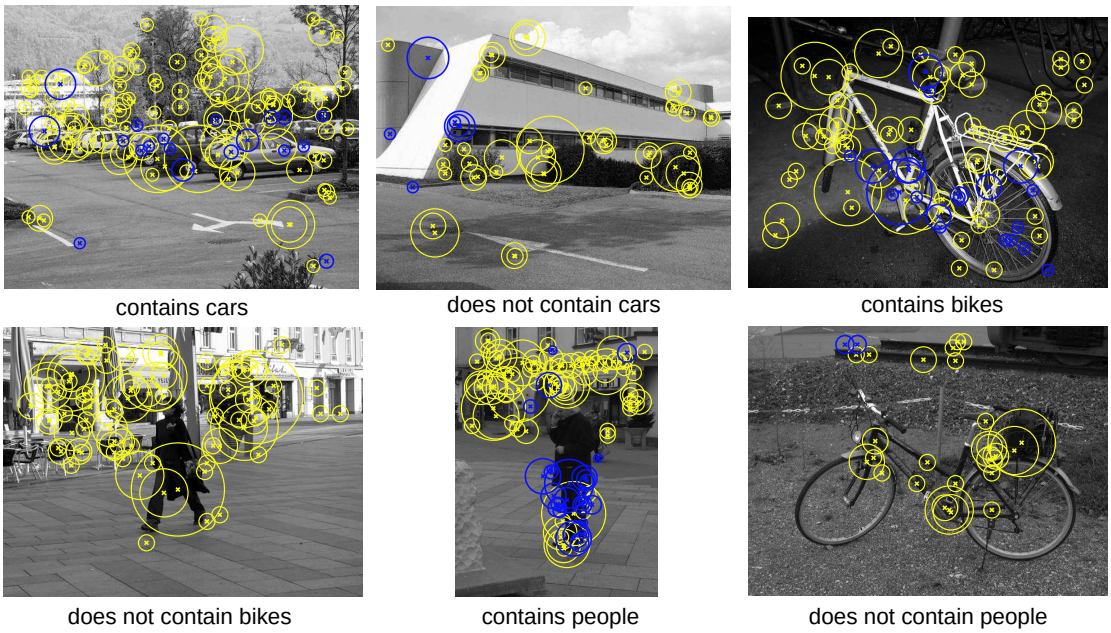


Figure 10: Correctly ranked test images along with the regions extracted by the Harris-Laplace (for cars and bikes) or Laplacian of Gaussian detector (for people). Dark blue circles represent regions that are more likely than not to belong to the object, while light yellow circles more probably belong to the background. The caption below each image indicates the correct classification.

Acknowledgements

We acknowledge the help of Nando de Freitas, Guillaume Bouchard, Hendrik Kück and Navneet Dalal, and the financial support of the European project LAVA and the PASCAL Network of Excellence.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, 2002.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Multiple instance learning with generalized support vector machines. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, 2002.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple instance learning. *Adv. Neural Inf. Process. Sys.*, 15, 2003.

[4] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A Statistical model for general contextual object recognition. In *Eur. Conf. Comp. Vision*, 2004.

[6] P. Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *Workshop AI Stats.*, 2003.

[7] G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.*, 95:957–970, 2000.

[8] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, November 1995.

[9] G. Dorkó and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *Int. Conf. Comp. Vision*, 2003.

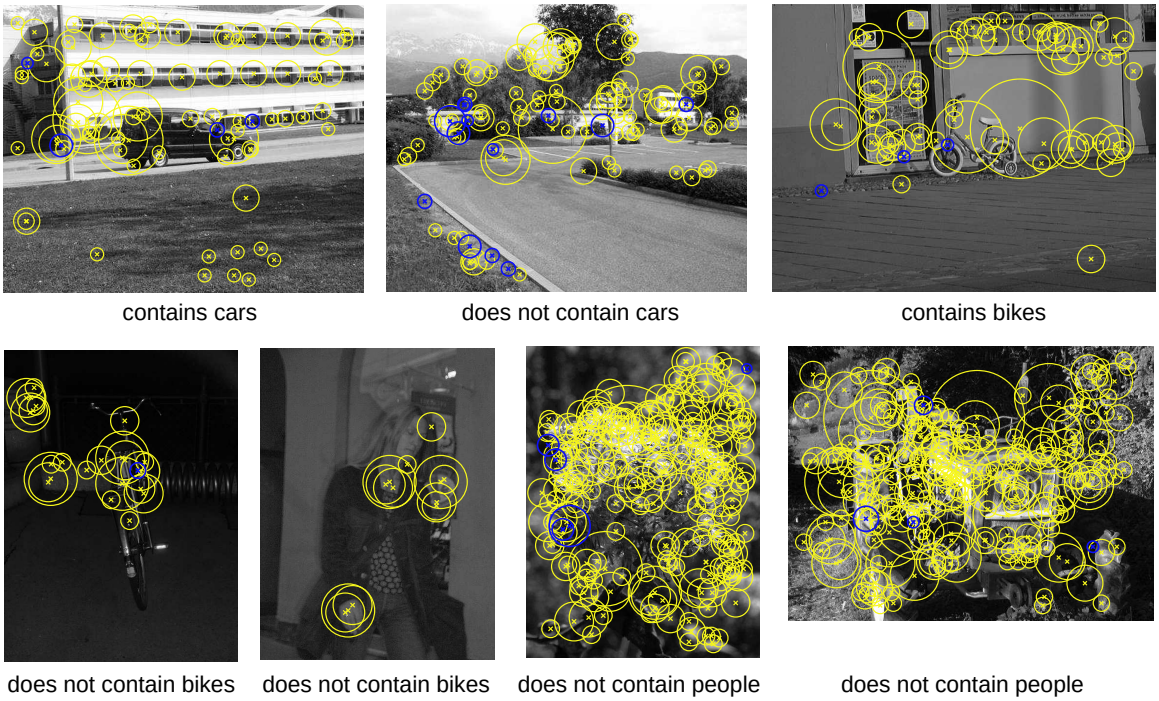


Figure 11: Incorrectly ranked test images along with the regions extracted by the Harris-Laplace (for cars and bikes) or Laplacian of Gaussian detector (for people). Dark blue circles represent regions that are more likely than not to belong to the object, while light yellow circles more probably belong to the background.

[10] P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Eur. Conf. Comp. Vision*, 2002.

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Conf. Comp. Vision and Pattern Recognit.*, 2003.

[12] V. Ferrari, T. Tuytelaars, and L. V. Gool. Simultaneous object recognition and segmentation by image exploration. In *Eur. Conf. Comp. Vision*, 2004.

[13] F. Jurie and C. Schmid. Scale-invariant shape features for recognition of object categories. In *Conf. Comp. Vision and Pattern Recognit.*, 2004.

[14] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 45(2):83–105, 2001.

[15] H. Kück, P. Carbonetto, and N. de Freitas. A Constrained semi-supervised learning approach to data association. In *Eur. Conf. Comp. Vision*, 2004.

[16] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2), 1998.

[17] J. S. Liu. Peskun's theorem and a modified discrete-stats Gibbs sampler. *Biometrika*, 83(3):681–682, 1996.

[18] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[19] J. S. Liu and Y. N. Wu. Parameter expansion for data augmentation. *J. Am. Stat. Assoc.*, 94(448):1264–1274, December 1999.

[20] T. Liu, A. W. Moore, and A. Gray. New Algorithms for efficient high-dimensional nonparametric classification. *Adv. Neural Inf. Process. Sys.*, 16, 2003.

[21] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[22] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. British Machine Vision Conf.*, 2002.

[23] D. McFadden. A Method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.

[24] K. Mikolajczyk and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Int. Conf. Comp. Vision*, 1999.

[25] K. Mikolajczyk and C. Schmid. An Affine invariant interest point detector. In *Eur. Conf. Comp. Vision*, 2002.

[26] K. Mikolajczyk and C. Schmid. A Performance evaluation of local descriptors. In *Conf. Comp. Vision and Pattern Recognit.*, 2003.

[27] T. Miller, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Faces and names in the news. In *Conf. Comp. Vision and Pattern Recognit.*, 2004.

[28] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: a graphical model relating features, objects, and scenes. *Adv. Neural Inf. Process. Sys.*, 16, 2003.

[29] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Eur. Conf. Comp. Vision*, 2004.

[30] S. Tham. *Markov Chain Monte Carlo for sparse Bayesian regression and classification*. PhD thesis, University of Melbourne, August 2002.

[31] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Research*, 1:211–244, 2001.

A Probability distributions

Here is a list of probability density functions used in this paper.

Name	**Density function**	**Parameters**
Bernoulli	$\mathcal{B}ern(x \mid \alpha) = \alpha^x(1-\alpha)^{(1-x)}$	$x \in \{0,1\}$, success rate $\alpha \in [0,1]$
Beta	$\mathcal{B}eta(x \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$x \in [0,1], a,b \geq 1$
Univariate Normal	$\mathcal{N}(x \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \times \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	$x \in \mathbb{R}$, mean μ, variance σ^2
Multivariate Normal	$\mathcal{N}(x \mid \mu,\Sigma) = \lvert 2\pi\Sigma\rvert^{-1/2} \times \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$	$x \in \mathbb{R}^F$, mean μ, $k \times k$ positive semi-definite covariance Σ
Inverse Gamma	$\mathcal{IG}(x \mid \alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}e^{-\beta/x}$	$x \in \mathbb{R}$, scale $a > 0$, shape $b > 0$

B Useful Identities

B.1 Matrix determinant

The determinant for a general block-partitioned matrix is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B|. \tag{B.1}$$

If a is a scalar and A is an $n \times n$ square matrix, then

$$|aA| = a^n|A|. \tag{B.2}$$

B.2 Matrix inversion

The matrix inversion formula for a symmetric matrix with block sub-matrices is

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} D^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}B^TA^{-1} & E^{-1} \end{bmatrix}, \tag{B.3}$$

where $D \triangleq A - BC^{-1}B^T$ and $E \triangleq C - B^TA^{-1}B$.

C Derivations

The following subsections provide the complete derivations of the posteriors used in Sec. 3.

C.1 Derivations of conditional posteriors for proposed classification model

We consider the model and notation introduced in Sec. 2. Following Bayes' rule, the conditional posterior for sampling the regresion coefficients is

$$\begin{aligned} p(\beta \,|\, \gamma, z) \;&\propto\; p(z \,|\, \gamma, \beta, x)\, p(\beta \,|\, \gamma, \delta^2) \\ &= \mathcal{N}\left(Q_\gamma \Psi_\gamma^T Z, Q_\gamma\right), \end{aligned} \tag{C.1}$$

where we define $Q \triangleq (\Psi^T\Psi + (\delta^2 S)^{-1})^{-1}$ and $Z \triangleq [z_1, z_2, \ldots, z_N]^T$.

Next, we derive the posterior of γ. It does not depend on β because it is combined with (C.1) to comprise the blocked move described in Sec. 3.1. Since we have to integrate out β, this derivation is somewhat involved. The conditional posterior of γ is proportional to the likelihood of z times the prior over γ. However, for the time being we can forget about the prior since it will end up cancelling out in the Metropolis-Hastings acceptance ratios (see Sec. 3.3). In any case, it is impractical to sample from $p(\gamma \,|\, \delta^2, z, x)$ directly, so we are really only interested in deriving the expression up to a constant term.

$$\begin{aligned} p(\gamma \,|\, \delta^2, z, x) \;&\propto\; p(z \,|\, \gamma, \delta^2, x) \\ &= \int p(z \,|\, \gamma, \beta, x)\, p(\beta \,|\, \gamma, \delta^2)\, d\beta \\ &= \int \left((2\pi)^{N+\Sigma\gamma} \left|\delta^2 S_\gamma\right|\right)^{-1/2} \exp\left(-\tfrac{1}{2}\left(Z^TZ - 2Z^T\Psi_\gamma\beta + \beta^T Q_\gamma^{-1}\beta\right)\right) d\beta \end{aligned} \tag{C.2a}$$

Temporarily defining $\Sigma \triangleq Q_\gamma$ and $\mu \triangleq \Sigma\Psi_\gamma^T Z$, and continuing from (C.2a), we get

$$\begin{aligned} p(z \,|\, \gamma, \delta^2, x) \;&=\; \int \left((2\pi)^{N+\Sigma\gamma} \left|\delta^2 S_\gamma\right|\right)^{-1/2} \exp\left(-\tfrac{1}{2}\left(Z^TZ + \beta^T\Sigma^{-1}\beta - 2\mu^T\Sigma^{-1}\beta + \mu^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\mu\right)\right) d\beta \\ &=\; \int \left((2\pi)^{N+\Sigma\gamma} \left|\delta^2 S_\gamma\right|\right)^{-1/2} \exp\left(-\tfrac{1}{2}\left(Z^TZ - \mu^T\Sigma^{-1}\mu\right)\right) \left((2\pi)^{\Sigma\gamma}|\Sigma|\right)^{1/2} \mathcal{N}(\beta \,|\, \mu, \Sigma)\, d\beta \\ &=\; \left((2\pi)^N \left|\delta^2 S_\gamma Q_\gamma^{-1}\right|\right)^{-1/2} \exp\left(-\tfrac{1}{2}\left(Z^TZ - Z^T\Psi_\gamma Q_\gamma \Psi_\gamma^T Z\right)\right) \\ &\propto\; \left|Q_\gamma(\delta^2 S_\gamma)^{-1}\right|^{1/2} \exp\left(-\tfrac{1}{2}Z^TZ\right) \exp\left(\tfrac{1}{2}Z^T\Psi_\gamma Q_\gamma \Psi_\gamma^T Z\right) \\ &\propto\; \sqrt{|Q_\gamma(\delta^2 S_\gamma)^{-1}| \exp\left(Z^T\Psi_\gamma Q_\gamma \Psi_\gamma^T Z\right)} \\ &=\; \sqrt{\left|\delta^2 S_\gamma \Psi_\gamma^T \Psi_\gamma + I_{\Sigma\gamma}\right|^{-1} \exp\left(Z^T\Psi_\gamma Q_\gamma \Psi_\gamma^T Z\right)}. \end{aligned} \tag{C.2b}$$

Finally, we consider the effect of activating and deactivating a kernel k on the prior $p(\gamma \,|\, a, b)$. We denote $\neq k$ to be shorthand for the set $\{k' \,|\, k \in \{1, 2, \ldots, K\}, k' \neq k\}$.

$$p(\gamma_k = 1, \gamma_{\neq k} \,|\, a, b) \;=\; \frac{p(\gamma_k = 1, \gamma_{\neq k} \,|\, a, b)}{p(\gamma_k = 1, \gamma_{\neq k} \,|\, a, b) + p(\gamma_k = 0, \gamma_{\neq k} \,|\, a, b)}$$

$$
\begin{aligned}
&= \frac{\Gamma\left(\Sigma\gamma_{\neq k}+1+a\right)\Gamma\left(K-\Sigma\gamma_{\neq k}-1+b\right)}{\Gamma\left(\Sigma\gamma_{\neq k}+1+a\right)\Gamma\left(K-\Sigma\gamma_{\neq k}-1+b\right)+\Gamma\left(\Sigma\gamma_{\neq k}+a\right)\Gamma\left(K-\Sigma\gamma_{\neq k}+b\right)} \\
&= \left(1+\frac{K-\Sigma\gamma_{\neq k}-1+b}{\Sigma\gamma_{\neq k}+a}\right)^{-1} \\
&= \frac{\Sigma\gamma_{\neq k}+a}{K+a+b-1}. \qquad \text{(C.3a)}
\end{aligned}
$$

The prior $p(\gamma \,|\, a, b)$ is derived in (2.7). The prior when proposing a flip from 1 to 0 is likewise

$$
\begin{aligned}
p(\gamma_k = 0, \gamma_{\neq k} \,|\, a, b) &= 1 - p(\gamma_k = 1 \,|\, \gamma_{\neq k} \,|\, a, b) \\
&= 1 - \frac{\sum \gamma_{\neq k}+a}{K+a+b-1} \\
&= \frac{K-\sum\gamma_{\neq k}+b-1}{K+a+b-1}. \qquad \text{(C.3b)}
\end{aligned}
$$

C.2 Derivation of parameter expansion algorithm for proposed classification model

We augment the sampling of the regression coefficients β by scaling the latent variables z using an expansion parameter α with an inverse Gamma prior, as in Sec. 3.2. Defining the transformation to be $t_\alpha(z_i) = z_i/\alpha$ and the Jacobian $J_\alpha(z) = \det\{\partial t_\alpha(z_i)/\partial z_i\} = \alpha^{-N}$, we sample α from the posterior

$$
\begin{aligned}
&p(\alpha \,|\, w, \gamma, \delta^2, \mu_\alpha, \nu_\alpha) \\
&\quad\propto p(t_\alpha(w) \,|\, \gamma, \delta^2, x)\, |J_\alpha(z)|\, p(\alpha \,|\, \mu_\alpha, \nu_\alpha) \\
&\quad= \int \alpha^{-N} p\big(\tfrac{w}{\alpha} \,\big|\, \gamma, \beta, x\big)\, p\big(\beta \,\big|\, \gamma, \delta^2, \tfrac{w}{\alpha}\big)\, p(\alpha \,|\, \mu_\alpha, \nu_\alpha)\, d\beta \\
&\quad= \int \alpha^{-N} \mathcal{N}\big(\tfrac{w}{\alpha} \,\big|\, \Psi_\gamma\beta, I_N\big)\, \mathcal{N}\big(\beta \,\big|\, \tfrac{1}{\alpha}\big(\Psi_\gamma^T\Psi_\gamma + (\delta^2 S_\gamma)^{-1}\big)^{-1}\Psi_\gamma^T W, \big(\Psi_\gamma^T\Psi_\gamma + (\delta^2 S_\gamma)^{-1}\big)^{-1}\big)\, \mathcal{IG}\big(\alpha^2 \,\big|\, \tfrac{\mu_\alpha}{2}, \tfrac{\nu_\alpha}{2}\big)\, d\beta \\
&\quad\propto \int \alpha^{-N-\mu_\alpha-2} \exp\Big(-\tfrac{1}{2}\Big(\tfrac{W^TW+\nu_\alpha}{\alpha^2} + \beta^T\big(2\Psi^T\Psi + (\delta^2 S)^{-1}\big)\beta - \tfrac{4}{\alpha}W^T\Psi\beta + \tfrac{1}{\alpha^2}W^T\Psi\big(\Psi^T\Psi + (\delta^2 S)^{-1}\big)^{-1}\Psi^T W \\
&\qquad\qquad \tfrac{4}{\alpha^2}W^T\Psi\left(2\Psi^T\Psi + (\delta^2 S)^{-1}\right)^{-1} - \tfrac{4}{\alpha^2}W^T\Psi\left(2\Psi^T\Psi + (\delta^2 S)^{-1}\right)^{-1}\Psi^T W\Big)\Big)\, d\beta \\
&\quad= \left(\alpha^2\right)^{-((N+\mu_\alpha+1)/2+1)} \exp\Big(-\tfrac{1}{2\alpha^2}\Big(\nu_\alpha + W^TW + W^T\Psi\Big(\big(\Psi^T\Psi + S^{-1}\big)^{-1} - 2\left(\Psi^T\Psi + (2\delta^2 S)^{-1}\right)^{-1}\Big)\Psi^T W\Big)\Big)\, \alpha \\
&\quad\propto \mathcal{IG}\Big(\alpha^2 \,\Big|\, \tfrac{N+\mu_\alpha+1}{2}, \tfrac{1}{2}\left(\nu_\alpha + W^TW + W^T\Psi\big(Q - 2\widehat{Q}\big)\Psi^T W\right)\Big) \times \frac{d\alpha^2}{d\alpha}. \qquad \text{(C.4)}
\end{aligned}
$$

where $w \triangleq t_{\alpha_0}^{-1}(z)$, $W \triangleq [w_1, w_2, \ldots, w_N]^T$, Q is defined as before and $\widehat{Q} \triangleq (\Psi^T\Psi + (2S)^{-1})^{-1}$. After the third line we remove the γ subscript from the variables Ψ, S and Q for clarity. Note that when S is set to the stabilised g-prior, $\widehat{Q} = (2\delta^2)((2\delta^2+1)\Psi^T\Psi + \epsilon I_K)^{-1}$.

C.3 Derivation of z likelihood with stabilised g-prior

We continue the derivation of the z likelihood, now with the coefficient scale prior set to the g-prior with the addition of a stabilising term, $S = (\Psi^T\Psi + \epsilon I_K)^{-1}$. In particular, we are interested an expressions for the likelihood when γ_k is set to 0 and when γ_k is set to 1.

First, we introduce some notation. We use the subscript 0 on the variables Ψ, S and Q and to be shorthand for $\{\gamma_k = 0, \gamma_{\neq k}\}$. We similarly define the subscript 1 to be $\{\gamma_k = 1, \gamma_{\neq k}\}$. Additionally, it will be useful to define Ψ_{γ_k} to be the column vector of the kernel response matrix introduced by $\gamma_k = 1$, $\Sigma\gamma_{\neq k} \triangleq \sum_{k'\neq k}\gamma_{k'}$, I_N to be the $N \times N$ identity matrix, $1_{N\times M}$ to be an $N \times M$ matrix of ones, $A_k \triangleq \Psi_0^T\Psi_{\gamma_k}$, $v \triangleq \frac{1+\delta^2}{\delta^2}$, and ζ_k and ξ_k to be the kth row, kth column elements in matrices S and Q, respectively.

ζ_k and ξ_k are analogous to the element E^{-1} in the matrix inversion formula (B.3), so the expressions for these two variables are

$$\begin{aligned}
\zeta_k &= \left(\Psi_{\gamma_k}^T \Psi_{\gamma_k} + \epsilon - A_k^T S_0 A_k\right)^{-1} \\
&= \left(\epsilon + \Psi_{\gamma_k}^T \left(I_N - \Psi_0 S_0 \Psi_0^T\right) \Psi_{\gamma_k}\right)^{-1} \\
\xi_k &= \left(v\Psi_{\gamma_k}^T \Psi_{\gamma_k} + \tfrac{\epsilon}{\delta^2} - v^2 A_k^T Q_0 A_k\right)^{-1} \\
&= \left(\tfrac{\epsilon}{\delta^2} + v\Psi_{\gamma_k}^T \left(I_N - v\Psi_0 Q_0 \Psi_0^T\right) \Psi_{\gamma_k}\right)^{-1}.
\end{aligned}$$

Following from the definitions in Sec. C.1 and the notation introduced here, we have

$$\begin{aligned}
Q_0 &= \left(\Psi_0^T \Psi_0 + (\delta^2 S_0)^{-1}\right)^{-1} \\
&= \left(\Psi_0^T \Psi_0 + (\delta^2)^{-1}(\Psi_0^T \Psi_0 + \epsilon I_{\Sigma\gamma})\right)^{-1} \\
&= \left(v\Psi_0^T \Psi_0 + \tfrac{\epsilon}{\delta^2} I_{\Sigma\gamma}\right)^{-1}.
\end{aligned}$$

When $\gamma_k = 0$, the expression follows straight from (C.2b):

$$p(z \,|\, \gamma_k{=}0, \gamma_{\neq k}, \delta^2, x) \;\propto \sqrt{(\delta^2)^{-\Sigma\gamma_{\neq k}} \left|Q_0 S_0^{-1}\right| \times \exp\left(Z^T \Psi_0 Q_0 \Psi_0^T Z\right)}. \tag{C.5a}$$

We need to do a bit more work in the derivation when $\gamma_k = 1$. We apply the formulas for matrix inversion (B.3) and for the matrix dterminant (B.1) to obtain

$$\begin{aligned}
p(z \,|\, \gamma_k{=}1, \gamma_{\neq k}, \delta^2, x) \;&\propto \sqrt{(\delta^2)^{-(\Sigma\gamma_{\neq k}+1)} \left|Q_1 S_1^{-1}\right| \exp\left(Z^T \Psi_1 Q_1 \Psi_1^T Z\right)} \\
&= \sqrt{(\delta^2)^{-(\Sigma\gamma_{\neq k}+1)}} \left| \frac{\Psi_1^T \Psi_1 + \epsilon I_{\Sigma\gamma_{\neq k}+1}}{v\Psi_1^T \Psi_1 + \frac{\epsilon}{\delta^2} I_{\Sigma\gamma_{\neq k}+1}} \right|^{1/2} \\
&\quad \times \exp\left(\tfrac{1}{2} Z^T \left(\begin{bmatrix} \Psi_0 & \Psi_{\gamma_k} \end{bmatrix} \begin{bmatrix} Q_0 + v^2 \xi_k Q_0 A_k A_k^T Q_0^T & -v\xi_k Q_0 A_k \\ -v\xi_k A_k^T Q_0^T & \xi_k \end{bmatrix} \begin{bmatrix} \Psi_0^T \\ \Psi_{\gamma_k}^T \end{bmatrix} \right) Z \right) \\
&= \sqrt{\xi_k |Q_0| \left(\zeta_k (\delta^2)^{\Sigma\gamma_{\neq k}+1} |S_0|\right)^{-1}} \\
&\quad \times \exp\Big(\tfrac{1}{2} Z^T \Big(\Psi_0 Q_0 \Psi_0^T + v^2 \xi_k \Psi_0 Q_0 A_k A_k^T Q_0^T \Psi_0^T - v\xi_k \Psi_0 Q_0 A_0 \Psi_{\gamma_k}^T \\
&\qquad\qquad - v\xi_k \Psi_{\gamma_k} A_k^T Q_0^T \Psi_0^T + \xi_k \Psi_{\gamma_k}^T \Psi_{\gamma_k} \Big) Z \Big) \\
&= \sqrt{\xi_k |Q_0| \left(\zeta_k (\delta^2)^{\Sigma\gamma_{\neq k}+1} |S_0|\right)^{-1}} \\
&\quad \times \exp\left(Z^T \Psi_0 Q_0 \Psi_0^T Z + \xi_k \left(Z^T \left(I_N - v\Psi_0 Q_0 \Psi_0^T\right) \Psi_{\gamma_k}\right)^2\right)^{1/2}.
\end{aligned} \tag{C.5b}$$

Since we are going to compute Metropolis-Hastings acceptance probabilities, we are interested in the ratios of the z likelihoods for the possible assignments of γ_k. The ratios simplify the expressions (C.5a) and (C.5b) somewhat:

$$\frac{p(z \,|\, \gamma_k{=}1, \gamma_{\neq k}, \delta^2, x)}{p(z \,|\, \gamma_k{=}0, \gamma_{\neq k}, \delta^2, x)} = \sqrt{\xi_k (\zeta_k \delta^2)^{-1} \times \exp\left(\xi_k \left(Z^T \left(I_N - v\Psi_0 Q_0 \Psi_0^T\right) \Psi_{\gamma_k}\right)^2\right)}. \tag{C.6}$$

The ratio $p(z \,|\, \gamma_k{=}0, \gamma_{\neq k}, \delta^2, x) / p(z \,|\, \gamma_k{=}1, \gamma_{\neq k}, \delta^2, x)$ follows similarly. Note that when we remove the stabilisation term ($\epsilon = 0$), the ratio (C.6) becomes

$$\frac{p(z \,|\, \gamma_k{=}1, \gamma_{\neq k}, \delta^2, x)}{p(z \,|\, \gamma_k{=}0, \gamma_{\neq k}, \delta^2, x)} = \sqrt{\frac{1}{1+\delta^2} \exp\left(\frac{\delta^2}{1+\delta^2} \times \frac{\left(Z^T \left(I_N - \Psi_0 S_0 \Psi_0^T\right) \Psi_{\gamma_k}\right)^2}{\Psi_{\gamma_k}^T \left(I_N - \Psi_0 S_0 \Psi_0^T\right) \Psi_{\gamma_k}}\right)},$$

which is the same result obtained by [30].

D Efficient algorithm for Metropolised Gibbs sampler

The algorithm METROPOLISED-GIBBS-SAMPLER-γ samples γ with time complexity $O(NK^2)$. We assume the algorithm is provided with computed values for the kernel response matrix Ψ, the Gram matrix $\Psi^T\Psi$ and the stabilised g-prior S. $\langle X \rangle$ means the value X inside the brackets has already been computed, and $\mathcal{U}_{[0,1]}$ denotes a random variate uniformly distributed on the interval $[0, 1]$. It is interesting to note that the algorithm fails when either ζ_k or ξ_k is negative, and these conditions imply either S or Q is not positive semi-definite. Therefore, this sampler ensures that β posterior is a proper distribution (provided the δ^2 samples are behaved).

METROPOLISED-GIBBS-SAMPLER-γ

$v \leftarrow \frac{1+\delta^2}{\delta^2}$
$Q \leftarrow \left(\langle \Psi^T\Psi \rangle + (\delta^2 S)^{-1}\right)^{-1}$
Compute $\langle \Psi^T Z \rangle$
for $k = 1$ **to** K **do** ▷ Preferably in a random order.
  if $\gamma_k = 0$ **then**
    if $\mathcal{U}_{[0,1]} < \frac{\Sigma\gamma+a}{K+a+b-1}$ **then**
      PROPOSE-GIBBS-BIRTH-γ
  else if $\mathcal{U}_{[0,1]} < \frac{K-\Sigma\gamma+b}{K+a+b-1}$ **then**
    PROPOSE-GIBBS-DEATH-γ

PROPOSE-GIBBS-BIRTH-γ

Compute Ψ_{γ_k}
Compute $\langle \Psi_{\gamma_k}^T \Psi_{\gamma_k} \rangle$
Compute $\langle \Psi^T \Psi_{\gamma_k} \rangle$
Compute $\langle \Psi_{\gamma_k}^T Z \rangle$
$\langle S\Psi^T\Psi_{\gamma_k} \rangle \leftarrow S\langle \Psi^T\Psi_{\gamma_k} \rangle$
$\langle Q\Psi^T\Psi_{\gamma_k} \rangle \leftarrow Q\langle \Psi^T\Psi_{\gamma_k} \rangle$
$\zeta_k \leftarrow 1/\left(\epsilon + \langle \Psi_{\gamma_k}^T\Psi_{\gamma_k} \rangle - \langle \Psi^T\Psi_{\gamma_k} \rangle^T \langle S\Psi^T\Psi_{\gamma_k} \rangle\right)$
$\xi_k \leftarrow 1/\left(\epsilon/\delta^2 + v\langle \Psi_{\gamma_k}^T\Psi_{\gamma_k} \rangle - v^2 \langle \Psi^T\Psi_{\gamma_k} \rangle^T \langle Q\Psi^T\Psi_{\gamma_k} \rangle\right)$
$x \leftarrow \langle \Psi_{\gamma_k}^T Z \rangle - v\langle \Psi^T Z \rangle^T \langle Q\Psi^T\Psi_{\gamma_k} \rangle$
if $\mathcal{U}_{[0,1]} < \min\left\{1, \sqrt{\frac{\xi_k}{\delta^2\zeta_k} \times \exp\left(\xi_k x^2\right)}\right\}$ **then**
  GIBBS-ACTIVATE-γ

PROPOSE-GIBBS-DEATH-γ

$\Psi_{\gamma_k} \leftarrow$ column k of Ψ
$\langle \Psi^T\Psi_{\gamma_k} \rangle \leftarrow$ column k of $\langle \Psi^T\Psi \rangle$ with all rows except row k
$\langle \Psi_{\gamma_k}^T Z \rangle \leftarrow$ row k of $\langle \Psi^T Z \rangle$
$\langle \Psi^T Z \rangle^{(\text{new})} \leftarrow$ all rows of $\langle \Psi^T Z \rangle$ except row k
$\zeta_k \leftarrow$ row k, column k entry of S
$V_k \leftarrow$ column k of S with all rows except row k
$S^{(\text{new})} \leftarrow$ all rows and columns of S except row k and column k
$S^{(\text{new})} \leftarrow S^{(\text{new})} - V_k V_k^T/\zeta_k$
$\xi_k \leftarrow$ row k, column k entry of Q
$V_k \leftarrow$ column k of Q with all rows except k
$Q^{(\text{new})} \leftarrow$ all rows and columns of Q except row k and column k
$Q^{(\text{new})} \leftarrow Q^{(\text{new})} - V_k V_k^T/\xi_k$
$x \leftarrow \langle \Psi_{\gamma_k}^T Z \rangle^{(\text{new})} - v\langle \Psi^T Z \rangle^T Q^{(\text{new})} \langle \Psi^T\Psi_{\gamma_k} \rangle$
if $\mathcal{U}_{[0,1]} < \min\left\{1, \sqrt{\frac{\delta^2\zeta_k}{\xi_k} \times \exp\left(-\xi_k x^2\right)}\right\}$ **then**
  GIBBS-DEACTIVATE-γ

GIBBS-ACTIVATE-γ

1 Ψ $\leftarrow [\Psi \ \Psi_{\gamma_k}]$

2 $\langle \Psi^T \Psi \rangle$ $\leftarrow \begin{bmatrix} \langle \Psi^T \Psi \rangle & \langle \Psi^T \Psi_{\gamma_k} \rangle \\ \langle \Psi^T \Psi_{\gamma_k} \rangle^T & \langle \Psi_{\gamma_k}^T \Psi_{\gamma_k} \rangle \end{bmatrix}$

3 $\langle \Psi^T Z \rangle$ $\leftarrow \begin{bmatrix} \langle \Psi^T Z \rangle \\ \langle \Psi_{\gamma_k}^T Z \rangle \end{bmatrix}$

4 S $\leftarrow \begin{bmatrix} S + \zeta_k \langle S\Psi^T \Psi_{\gamma_k} \rangle \langle S\Psi^T \Psi_{\gamma_k} \rangle^T & -\zeta_k \langle S\Psi^T \Psi_{\gamma_k} \rangle \\ -\zeta_k \langle S\Psi^T \Psi_{\gamma_k} \rangle^T & \zeta_k \end{bmatrix}$

5 Q $\leftarrow \begin{bmatrix} Q + v^2 \xi_k \langle Q\Psi^T \Psi_{\gamma_k} \rangle \langle Q\Psi^T \Psi_{\gamma_k} \rangle^T & -v\xi_k \langle Q\Psi^T \Psi_{\gamma_k} \rangle \\ -v\xi_k \langle Q\Psi^T \Psi_{\gamma_k} \rangle^T & \xi_k \end{bmatrix}$

GIBBS-DEACTIVATE-γ

1 Ψ $\leftarrow$ all columns of Ψ except k

2 $\langle \Psi^T \Psi \rangle$ $\leftarrow$ all rows/columns of $\langle \Psi^T \Psi \rangle$ except row/column k

3 $\langle \Psi^T Z \rangle$ $\leftarrow \langle \Psi^T Z \rangle^{(\text{new})}$

4 S $\leftarrow S^{(\text{new})}$

5 Q $\leftarrow Q^{(\text{new})}$